Data Centers to Offer Ancillary Services

Mahdi Ghamkhari and Hamed Mohsenian-Rad
Department of Electrical Engineering, University of California at Riverside, Riverside, CA, USA
e-mails: {ghamkhari, hamed}@ee.ucr.edu

***Abstract*— Considering the growing number of Internet and cloud computing data centers being built in recent years and given the data centers' *major* and yet *flexible* electric load, they can be good candidates to offer ancillary services, such as voluntary load reduction, to a smart grid. In this paper, we investigate such potential within an analytical profit maximization framework to determine whether participation in an ancillary service market can be beneficial to data centers. The profit model that we introduce includes elements with respect to a) the data center's revenue obtained from the Internet services that the data center offers based on its service-level agreements (SLA), b) the data center's cost of electricity based on time-of-use prices, and c) the monetary compensation that the data center may receive due to offering ancillary services based on the existing ancillary service market models in the ERCOT (Electric Reliability Council of Texas) Independent System Operator. Our simulation results show that data centers can noticeably increase their profit by participating in voluntary load reduction. Their participation can also help the grid better maintain service quality and reliability.**

***Keywords*: Data center, ancillary service, voluntary load reduction, profit maximization, load resource, service-level agreement.**

I. Introduction

In an interconnected power system, Independent System Operators (ISOs) are responsible for coordinating, controlling, and monitoring the operation of the power grid for generation, transmission and distribution. An ISO is also responsible for maintaining a required level of power quality and reliability, e.g. by making a constant balance between supply and demand across the power grid. For this purpose, the ISOs rely on receiving different types of ancillary services from a variety of entities that are involved in the power system. Examples of ancillary services include frequency and voltage regulation, spinning reserve, and non-spinning reserve [1].

Ancillary services are usually procured from generators that are online and can increase or decrease their generation in response to the requests sent by ISOs. However, there is a growing interest towards procuring ancillary services not only from generators but also from load resources. Reduced transmission and distribution losses, increased transmission capacity and increased margin to voltage collapse are among the benefits in supplying ancillary services from load resources [2]. Load resources are paid to offer *voluntary load reduction*, with a compensation value equivalent to what a generator is paid for generating the same amount of electricity [3].

The ancillary service market model for controllable load in ERCOT ISO is shown in Fig. 1. A key entity to facilitate load participation in ancillary service market is Qualified Scheduling Entity (QSE). It acts as an interface between ERCOT and controllable load resources (CLRs). The QSE coordinates the operation of CLRs based on the commands it receives from ERCOT. The QSE also aggregates the ancillary services that CLRs may offer. The current total CLR capacity in ERCOT is about 36 MW [5]. They are eligible to participate in both *regulation* and *voluntary load reduction* services [4]. However, our focus in this paper is on the latter.

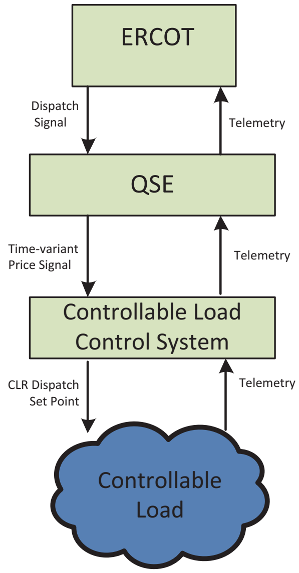

Fig. 1. The ancillary service market model for controllable load in ERCOT [4]. Data centers can register as load resources to offer voluntary load reduction.

In this paper, we would like to investigate the potential for Internet and cloud computing data centers to participate in an ancillary service market to offer voluntary load reduction. Data centers have two important features that make them good candidates to offer such services. First, data centers, such as those built and operated by Google, Microsoft, and Amazon, have significant daily and peak power consumption. For example, the peak power load of Microsoft's data center in Quincy, WA is 48 megawatts, which is enough to power 40,000 homes [6]. Second, data centers have flexible load and they are able to respond to the QSE's signals quickly and reduce their power consumption by switching off a group of computer servers or by migrating a portion of their workload to another data center [7], [8]. These features suggest that data center participation in the ancillary service market can be promising. To the best of our knowledge, this paper is the first to study the capability of data centers in offering ancillary services, in particular in form of voluntary load reduction. Our contributions in this paper can be summarized as follows.

- We develop a mathematical model for data center's profit when it offers ancillary services. Our model includes elements with respect to the data center's *revenue* ob-

tained from the Internet services it offers, the data center's *cost* of electricity, and the *compensation* that the data center receives for offering ancillary services. We take into account server's power consumption profiles, data center's power usage effectiveness, price of electricity, workload statistics, and service-level agreements (SLAs).
- We propose an optimization-based *profit maximization strategy* for data centers, when they offer ancillary services in form of voluntary load reduction. To gain insights, we also provide a *geometric interpretation* of the optimal solution of the profit maximization problem.
- Using experimental data, e.g., for workload, price of electricity, and SLA parameters, we assess the performance of the proposed optimization-based profit maximization strategy via computer simulations. We show that a data center can noticeably increase its profit by participating in a voluntary load reduction ancillary service program.

The rest of this paper is organized as follows. The system model is described in Section II. The mathematical expressions for revenue, cost, and ancillary service compensation are derived in Section III. Our proposed profit maximization design framework is discussed in Section IV. Simulation results are presented in Section V. The paper is concluded in Section VI.

II. SYSTEM MODEL

A. Power Consumption

Consider an Internet or cloud computing data center with $M_{\max}$ computer servers. The total power consumption in a data center is obtained by adding the total power consumption at computer servers to the total power consumption at the facility, e.g., for cooling, lighting, etc. For a data center, *power usage effectiveness* (PUE), denoted by E_{usage}, is defined as the ratio of the data center's total power consumption to the data center's power consumption at the computer servers [9]. The PUE is considered as a measure for data center's energy efficiency. Currently, the typical value for most data centers is around 2.0. However, recent studies have suggested that many data centers can soon reach a PUE of 1.7. A few state-of-the art facilities have reached a PUE of 1.2 [9].

Let P_{idle} denote the average idle power draw of a single server and P_{peak} denote the average peak power when a server is handling a service request. The ratio $P_{\text{peak}}/P_{\text{idle}}$ denotes the power elasticity of servers. Higher elasticity means less power consumption when the server is idle, not handling any service request. Let $M \leq M_{max}$ denote the number of servers that are 'on' at data center. The total electric power consumption associated with the data center can be obtained as [10]:

$$P = M[P_{\text{idle}} + (E_{\text{usage}} - 1)P_{\text{peak}} + \\ (P_{\text{peak}} - P_{\text{idle}})U], \tag{1}$$

where U is the CPU utilization of servers. From (1), the power consumption at data center increases as we turn on more computer servers or run servers at higher utilization.

B. Electricity Price

The electricity pricing models that are deployed for each region usually depend of whether the electricity market is regulated or deregulated in that region. In ERCOT, the electricity market is mostly deregulated. Therefore, the prices may vary during the day due to the fluctuations in the wholesale market. Some of the common non-flat retail pricing tariffs in deregulated electricity markets include: Day-ahead pricing (DAP), time-of-use pricing (TOUP), critical-peak pricing (CPP), and real-time pricing (RTP). In our system model, the instantaneous price of electricity is denoted by ω which is assumed to be known at least 15 minutes in advance.

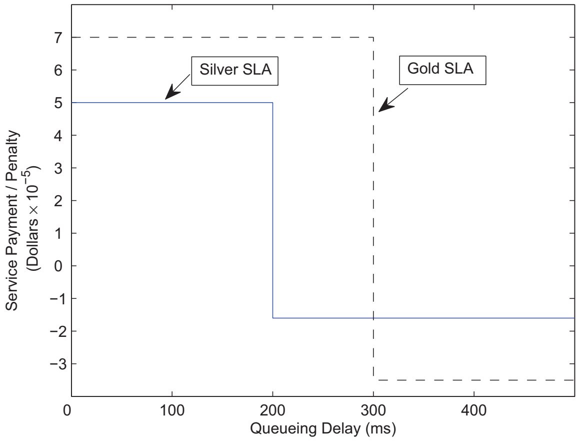

Fig. 2. Two sample service-level agreements (SLAs) in data centers [11].

C. Voluntary Load Reduction as Ancillary Service

The ERCOT ancillary service market is designed with a number of features to reward consumers that are willing to curtail their load when needed to help maintain system reliability. In particular, consumers that offer voluntary load reduction are compensated in dollars per megawatt hour of load reduction at rates that are set based on the market clearing prices within the ERCOT-operated markets [3]. Given the data centers' major and flexible load, they can register as load resources and respond to the load reduction request signals that are sent by the QSEs in their regions. The exact type of contracts to compensate such data centers would depend on the mutual agreements between the QSEs and the data centers [3]. Here, we assume that the QSE may send out a load reduction request periodically, e.g., once every 15 minutes. The request contains a compensation function $\tau(\cdot)$, which is calculated by the QSE based on a market clearing price analysis and indicates the dollars to be paid to the data center for each megawatt hour voluntary load reduction. The compensation function may or may not be linear. More details on compensation function will be discussed in Section III-C.

D. Quality-of-Service

Because of the limited computational capacity of data centers and given the stochastic nature of most practical workload, data centers cannot process the incoming service requests immediately after they arrive. Therefore, all arriving service requests are first placed in a queue until they can be handled by an available server. In order to satisfy quality-of-service requirements, the waiting time/queuing delay for each incoming service request should be limited within a certain range which is determined by the *Service Level Agreement* (SLA). The exact SLA depends on the type of service offered which may range from cloud-based computational tasks to video streaming and HTML web services. Examples of two typical SLAs are shown in Fig. 2 [11]. In this figure, each SLA is identified by three non-negative parameters D, δ, and γ. Parameter D indicates the maximum waiting time that a service request can tolerate. Parameter δ indicates the service money that the data center receives when it handles a single service request before deadline D. Parameter γ indicates the penalty that the data center has to pay every time it *cannot* handle a service request before deadline D. For the Gold SLA in Fig. 2, we have $D = 300$ ms, $\delta = 7 \times 10^{-5}$ dollars, and $\gamma = 3.5 \times 10^{-5}$ dollars. For the Silver SLA, we have $D = 200$ ms, $\delta = 5 \times 10^{-5}$ dollars, and $\gamma = 1.6 \times 10^{-5}$ dollars.

E. Service Rate

Let μ denote the rate at which service requests are removed from the queue and handled by a server. The service rate depends on the number of servers that are switched on. Let S denote the time it takes for a server to finish handling a service request. Each server can handle $\kappa = 1/S$ service requests per second. Therefore, the total service rate is obtained as

$$\mu = \kappa M \quad \Rightarrow \quad M = \frac{\mu}{\kappa}. \tag{2}$$

As we increase the number of switched on servers and thus the service rate, more service requests can be handled before the SLA-deadline D, which in turn increases the payments that the data center receives. However, it also increases the data center's power consumption and thus the data center's energy expenditure. Furthermore, turning on more computer servers degrades the data center's ability to offer load reduction as an ancillary service. Therefore, there is a *trade-off* in selecting the data center's service rate, as we will discuss next.

III. Revenue, Cost, and Compensation Models

The rate at which *service requests* arrive at a data center can vary over time. To improve data center's performance, the number of switched on servers M should be adjusted in proportion to demand. More servers should be turned on when service requests are received at higher rates. However, because of the tear-and-wear cost of switching servers on and off, and also due to the delay in changing the status of a computer, M cannot be changed rapidly. It is rather desired to be updated only every few minutes. Therefore, we divide running time of data center into a sequence of time slots $\Lambda_1, \Lambda_2, \cdots, \Lambda_N$, each one with length T. The number of switched on servers are updated only at the beginning of each time slot. We assume that $T = 15$ minutes, so that making decision about the number of switched on computers is periodic and is done at the same time that the data center receives a load reduction request from the QSE. For the rest of this section, we focus on mathematically modeling the energy cost, Internet service revenue, and ancillary service compensation for a data center at each time slot $\Lambda \in \Lambda_1, \Lambda_2, \cdots, \Lambda_N$ as a function of service rate μ and consequently as a function of M, based on (2).

A. Revenue Modeling

Let $q(\mu)$ denote the probability that the waiting time for a service request exceeds the SLA-deadline D. Obtaining an analytical model for $q(\mu)$ requires a queueing theoretic analysis that we will provide in Section III-D. Next, assume that λ denotes the average rate of receiving service requests within time slot Λ of length T. The total revenue collected by the data center at the time slot of interest can be calculated as

$$Revenue = (1 - q(\mu))\delta\lambda T - q(\mu)\gamma\lambda T, \tag{3}$$

where $(1 - q(\mu))\delta\lambda T$ denotes the total payment received by the data center within interval T, for the service requests that are handled before the SLA-deadline, while $q(\mu)\gamma\lambda T$ denotes the total penalty paid by the data center within interval T for the service requests that are not handled before the SLA-deadline.

B. Cost Modeling

Within time interval T, each turned on server handles

$$\frac{T(1 - q(\mu))\lambda}{M} \tag{4}$$

service requests. This makes each server busy for $T(1 - q(\mu))\lambda/\kappa M$ seconds. By dividing the total CPU busy time by T, the CPU utilization for each server is obtained as

$$U = \frac{(1 - q(\mu))\lambda}{\kappa M}. \tag{5}$$

Replacing (2) and (5) in (1), the power consumption associated with the data center at the time slot of interest is obtained as

$$P(\mu) = \frac{a\mu + b\lambda(1 - q(\mu))}{\kappa}, \tag{6}$$

where $a \triangleq P_{\text{idle}} + (E_{\text{usage}} - 1)P_{\text{peak}}$ and $b \triangleq P_{\text{peak}} - P_{\text{idle}}$. Multiplying (6) by the electricity price ω, the total energy cost at the time interval of interest is obtained as

$$Cost = T\omega \left[\frac{a\mu + b\lambda(1 - q(\mu))}{\kappa}\right]. \tag{7}$$

C. Ancillary Service Compensation Model

Let Λ_0 denote the time slot right before the current time slot Λ. Also let P_0 denote the total power consumption at time slot Λ_0. If the data center reduces its load by $\Delta P = P_0 - P(\mu)$ compared to the previous time slot, then the QSE pays

$$Compensation = \tau(\Delta P) \tag{8}$$

to the data center in order to compensate for the voluntary load reduction ancillary service that is offered by the data center. The choice of compensation function $\tau(\cdot)$ is set by the QSE. If no load reduction ancillary service is needed at a time slot, then we simply have $\tau(\Delta P) = 0$. If the compensation function is linear, then we have $\tau(\Delta P) = c\Delta P$, where higher c indicates higher compensation rates, e.g., due to a more severe need for load reduction. Other forms of compensation functions may include quadratic or piece-wise linear functions. It is worth mentioning that once the compensation function is announced by the QSE, it is up to the data center to decide whether offering load reduction ancillary service is beneficial.

D. Probability Model of $q(\mu)$

Consider a new service request that arrives within time slot Λ. Let Q denote the number of service requests waiting in the service queue right before the arrival of the new service request. Since the data center's service rate is μ, it takes Q/μ seconds until all existing requests are removed from the queue. Hence, the new service request can be handled after Q/μ seconds since its arrival. According to the SLA, if $Q/\mu \leq D$, then the request is handled before the deadline D. If $Q/\mu > D$, the request is not handled before the deadline D and it is dropped. Therefore, we can model the SLA-deadline by a *finite-size queue* with the length $\mu\, D$. A service request can be handled before the SLA-deadline, if and only if it enters the aforementioned finite size queue. We assume that the service request rate has an arbitrary and *general* probability distribution function. On the other hand, since the service rate μ is fixed over each time interval of length T, $q(\mu)$ can be modeled as the *loss probability* of a G/D/1 queue. Therefore, following the queuing theoretic analysis in [12], we can obtain

$$q(\mu) = \alpha(\mu)\, e^{-\frac{1}{2} \min\limits_{n \geq 1} \; m_n(\mu)}, \tag{9}$$

where

$$\alpha(\mu) = \frac{1}{\lambda\sqrt{2\pi}\sigma} e^{\frac{(\mu-\lambda)^2}{2\sigma^2}} \int\limits_{\mu}^{\infty} (r-\mu) e^{-\frac{(r-\lambda)^2}{2\sigma^2}} dr \tag{10}$$

and for each $n \geq 1$ we have

$$m_n(\mu) = \frac{(D\mu + n(\mu-\lambda))^2}{nC_\lambda(0) + 2\sum\limits_{l=1}^{n-1} C_\lambda(l)(n-l)}. \tag{11}$$

Here, $\sigma = C_\lambda(0)$ and C_λ denotes the auto-covariance of the service request rate [13].

IV. Profit Maximization

A. The Case without Offering Ancillary Service

For the case where the data center does not offer ancillary service, its profit at each time slot Λ is obtained as

$$Profit = Revenue - Cost, \tag{12}$$

where revenue is as in (3) and cost is as in (7). We seek to choose the data center's service rate μ to maximize profit. This can be expressed as the following optimization problem:

$$\begin{aligned} \underset{\lambda \leq \mu \leq \kappa M_{max}}{\textbf{Maximize}} \;\; & T\lambda\left[(1-q(\mu))\delta - q(\mu)\gamma\right] - \\ & T\omega\left(\frac{a\mu + b\lambda(1-q(\mu))}{\kappa}\right), \end{aligned} \tag{13}$$

where the probability $q(\mu)$ is as in (9). We note that the service rate μ is lower bounded by λ. This is necessary to assure stabilizing the service request queue [12], [14]. We also note that problem (15) needs to be solved separately for every time slot $\Lambda \in \{\Lambda_1, \ldots, \Lambda_N\}$, i.e., once every T minutes.

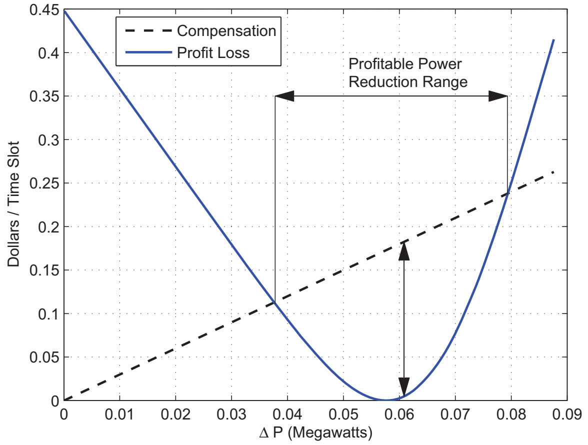

Fig. 3. An example to compare profit loss due to load reduction versus the compensation received due to offering ancillary service. The arrow indicates that the optimal amount of load reduction ΔP = 0.0608 Megawatts.

B. The Case with Offering Ancillary Service

For the case where data center does offer ancillary service, the data center's profit at each time slot Λ is obtained as

$$Profit = Revenue - Cost + Compensation, \tag{14}$$

where compensation is as in (8). We seek to choose the data center's service rate μ to maximize profit. This can be expressed as the following optimization problem:

$$\begin{aligned} \underset{\lambda \leq \mu \leq \kappa M_{max}}{\textbf{Maximize}} \;\; & T\lambda\left[(1-q(\mu))\delta - q(\mu)\gamma\right] - \\ & T\omega\left(\frac{a\mu + b\lambda(1-q(\mu))}{\kappa}\right) + \\ & T\tau\left(P_0 - \frac{a\mu + b\lambda(1-q(\mu))}{\kappa}\right), \end{aligned} \tag{15}$$

where the last term is based on the definition of ΔP in Section III-C and the expression for power consumption in (6).

C. Geometric Interpretation

Let $Profit_{Base}$ denote the optimal objective value of problem (13). That is, the maximum profit that a data center can obtain, by properly selecting its service rate, when the data center does *not* offer any ancillary service. Clearly, offering ancillary service is beneficial if the profit when offering ancillary service is greater than $Profit_{Base}$. That is,

$$\begin{aligned} & T\lambda\left[(1-q(\mu))\delta - q(\mu)\gamma\right] - \\ & T\omega\left(\frac{a\mu + b\lambda(1-q(\mu))}{\kappa}\right) + \\ & T\tau\left(P_0 - \frac{a\mu + b\lambda(1-q(\mu))}{\kappa}\right) \geq Profit_{Base}. \end{aligned} \tag{16}$$

From the definition of ΔP, we have

$$P(\mu) = P_0 - \Delta P. \tag{17}$$

Therefore, we can write μ in terms of ΔP as follows:

$$\mu = P^{-1}(P_0 - \Delta P), \tag{18}$$

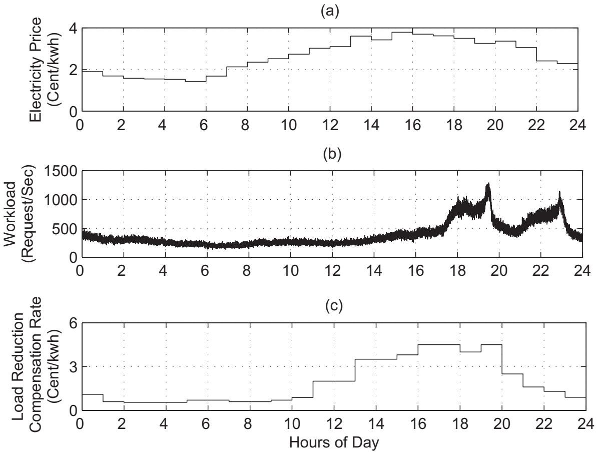

Fig. 4. A sample set of data over 24 hours that we used in our simulations. (a) Time-of-use electricity prices [15]. (b) Internet workload [16]. (c) The price of compensating voluntary load reduction as ancillary service [3].

where $P^{-1}(\cdot)$ denotes the inverse of function $P(\mu)$ in (6). Note that, since $q(\mu)$ is a non-increasing function of μ [13, Theorem 1], $P(\mu)$ is an increasing function of service rate μ. Therefore, $P(\mu)$ is an invertible function. From (18) and after reordering the terms, we can rewrite condition (16) as

$$\tau(\Delta P) > Profit_{Base}/T - \lambda \left[(1 - q(P^{-1}(P_0 - \Delta P)))\delta \right. \\ \left. - q(P^{-1}(P_0 - \Delta P))\gamma\right] + \omega(P_0 - \Delta P). \tag{19}$$

First, we note that both sides of condition (19) are written in terms of ΔP as the only variable. Second, while the left hand side in (19) is the ancillary service compensation function at load reduction level ΔP, the right hand side of (19) denotes the data center's *profit loss* due to reducing its load by ΔP if the QSE does *not* compensate the offered load reduction service. Therefore, we can conclude that offering voluntary load reduction at level ΔP is profitable if and only if

$$Compensation > Profit\,Loss, \tag{20}$$

where $Profit\,Loss$ is defined as the expression in the right hand side of (19). The geometric interpretation of the above condition is shown in Fig. 3. Here, the compensation function $\tau(\cdot)$ is assumed to be linear. We can see that offering load reduction ancillary service less than $\Delta P = 0.0378$ Megawatts or more than $\Delta P = 0.0793$ Megawatts is not profitable for the data center. Furthermore, we can see that the maximum profit is gained if the data center offers load reduction at $\Delta P = 0.0608$ Megawatts. This is essentially the same amount that the data center chooses to offer after solving problem (15).

V. Simulation Results

Consider a data center with a maximum of $M_{\max} = 50,000$ servers. The exact number of switched on servers M is updated periodically for each time slot of length $T = 15$ minutes by solving the profit maximization problems explained in Section IV. For each switched on server, we have P_{peak} = 200 watts and P_{idle} = 100 watts [7]. We assume that E_{usage} = 1.2 [9]. The electricity price information is based on the hourly real-time pricing tariffs currently practiced in Illinois Zone I, spanning from June 10, 2011 to July 9, 2011 [15]. We assume that $\kappa = 0.1$ and the Gold SLA is used. To simulate the total workload, we use the publicly available World Cup 98 web hits data, spanning from June 10, 1998 to July 9, 1998, as the trend for the incoming service requests [16]. We assume that the ancillary service compensation function is linear and is set according to the prices used in ERCOT [3]. A sample daily data set that we used in our simulation is shown in Fig. 4.

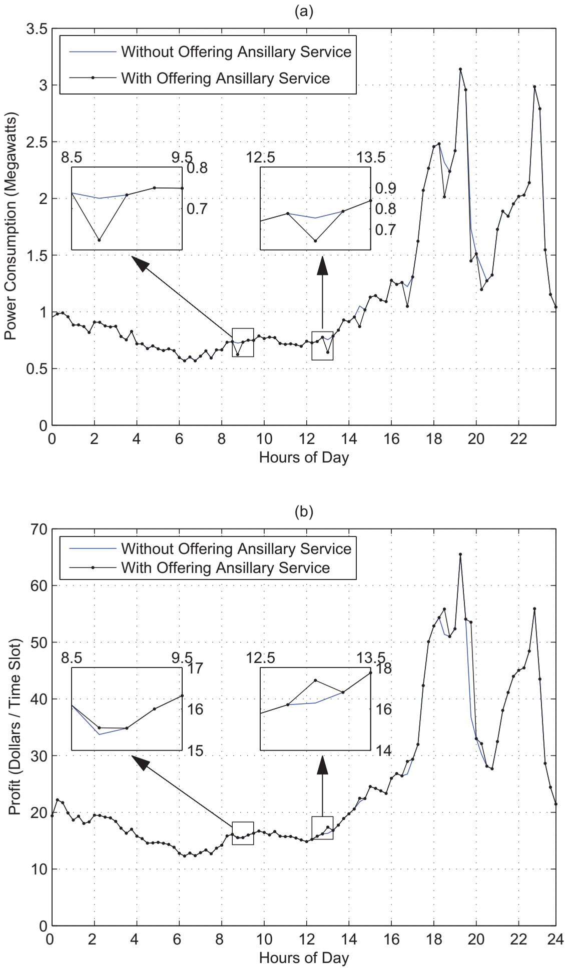

Fig. 5. The power consumption and profit in an example 24 hours running time of data center. Load reduction requests are received in seven time slots.

The results for a sample daily power consumption trend are shown in Fig. 5(a). It is assume that the QSE sends requests for load reduction in seven time slots. In all cases, the data center chooses to respond by reducing its load. Recall that the amount of load reduction is set based on the optimal solution of problem (15). The reduced power consumption for two time slots, one around 8:45 AM and one around 1:00 PM are zoomed in. The data center's corresponding profit is shown in 5(b). We can see that every time that the data center reduces its load in response to the QSE's request, its profit increases.

The daily increase in the data center's profit due to offering voluntary load reduction is shown in Fig. 6 over 30 days. On average, the data center's daily profit increases by 22.7 dollars, summing up to a monthly increase of 681 dollars. For the results in this figure, we have assumed that the QSE sends load reduction requests in 20% of the time slots. Clearly, a higher number of load reduction requests to be sent by the QSE can provide the data center with more opportunities to further increase its profit. This is shown in Fig. 7. The results in this figure are based on repeating the simulations in Fig. 6 for different number of time slots in which the QSE sends load reduction requests to load resources. We can see that, the data center's profit increases as the percentage of time slots with voluntary load reduction opportunity increases. If a load reduction request is sent by the QSE in every time slot, then the data center's profit can increase up to $\$3,287$ per month.

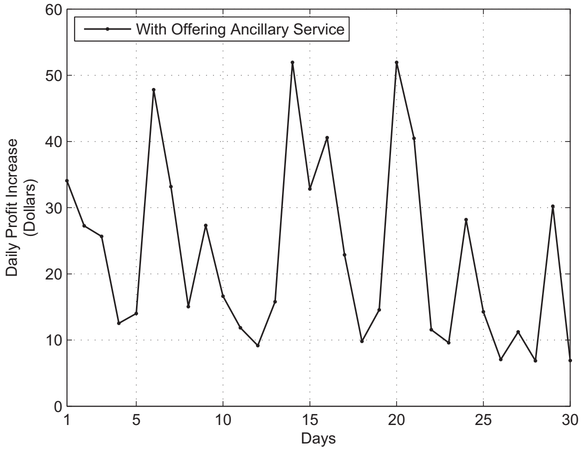

Fig. 6. The daily additional profit the data center gains over 30 days due to offering voluntary load reduction as ancillary service to smart grid.

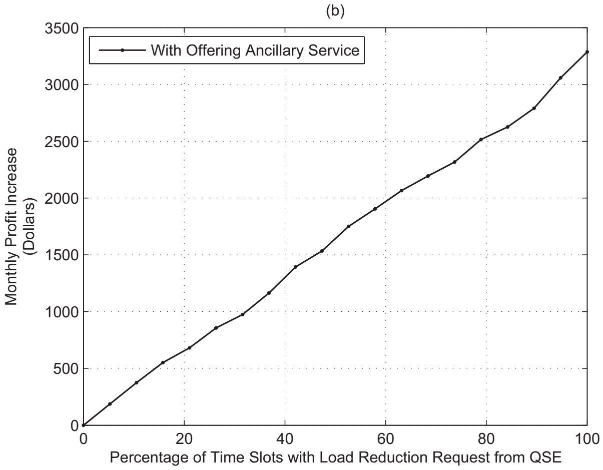

Fig. 7. The monthly additional profit the data center gains versus the percentage of time slots in which the QSE sends out load reduction requests.

VI. Conclusions and Future Work

This paper represents the first step towards enabling Internet and cloud computing data centers to offer ancillary services to smart grid. We particularly focused on the scenario where a data center is registered as a load resource to offer voluntary load reduction ancillary service. We proposed an analytical profit maximization framework for the data center to set its service rate and the amount of load reduction ancillary service at their optimal levels. Our profit model includes elements with respect to the data center's Internet service revenue, the cost of electricity, and the compensation it may receive to offer ancillary services. The model takes into account server's power consumption profiles, data center's power usage effectiveness, price of electricity, workload statistics, and SLAs. Our simulation results show that data centers can increase their profit by offering voluntary load reduction as ancillary service.

The results in this paper can be extended in several directions. First, while our focus in this paper is on voluntary load reduction, it is interesting to examine offering other forms of ancillary services such as frequency regulation. Second, other contract models between the data center and the QSE can be considered. In particular, one can extend the analysis such that the data center can submit bids to the ancillary service market. Finally, in addition to adjusting the operation of each data center using the proposed optimization-based approach, a group of data centers can coordinate their operation to further increase their profit. In particular, the idea of migrating a portion of service workload from one data center to another which has already been studied in the literature for reducing data centers cost of electricity can be extended to the scenario where data centers benefit from offering ancillary services.

Acknowledgment

The authors appreciate the helpful discussions with Paul Wattles, the Supervisor for Demand Response at ERCOT.

References

[1] M. Shahidehpour, H. Yamin, and Z. Li, *Market Operations in Electric Power Systems*. New York, NY: IEEE Press, 2002.

[2] J. D. Kueck, A. F. Snyder, F. Li, and I. B. Snyder, "Use of responsive load to supply ancillary services in the smart grid: Challenges and approach," in *Proc. of IEEE Smart Grid Comm*, Oct. 2010.

[3] ERCOT, "Load participation in the ercot nodal market," www.ercot.com.

[4] ——, "Controllable load resource (CLR) participation in the ERCOT market," www.ercot.com.

[5] P. Wattles, "Load resources providing ancillary services in electric reliability council of texas (ERCOT)," in *Proc. of IEEE Conference on Innovative Smart Grid Technologies*, Washington, DC, jan 2012.

[6] R. Katz, "Tech titans building boom," *IEEE Spectrum*, vol. 46, no. 2, pp. 40–54, Feb. 2009.

[7] M. Ghamkhari and H. Mohsenian-Rad, "Optimal integration of renewable energy resources in data centers with behind-the-meter renewable generators," in *Proc. of the IEEE International Conference on Communications*, Ottawa, Canada, June 2012.

[8] H. Mohsenian-Rad and A. Leon-Garcia, "Coordination of cloud computing and smart power grids," in *Proc. of the IEEE Smart Grid Communications Conference*, Gaithersburg, MD, oct 2010.

[9] United States Environmental Protection Agency, "EPA report on server and data center energy efficiency," Final Report to Congress, Aug. 2007.

[10] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in *Proc. of the ACM SIGCOMM*, Barcelona, Spain, 2009.

[11] D. Kusic, J. O. Kephart, J. E. Hanson, N. Kandasamy, and G. Jiang, "Power and performance management of virtualized computing environments via lookahead control," in *Proc. of IEEE International Conference on Autonomic Computing*, Nice, France, June 2008.

[12] H. S. Kim and N. B. Shroff, "Loss probability calculations and asymptotic analysis for finite buffer multiplexers," *IEEE/ACM Transactions on Networking*, vol. 9, no. 6, pp. 755 –768, dec 2001.

[13] M. Ghamkhari and H. Mohsenian-Rad, "Energy and performance management of green data centers: A profit maximization approach," *submitted to IEEE Transactions on Smart Grid, 2012*.

[14] Z. Liu, M. Lin, A. Wierman, S. Low, and L. L. H. Andrew, "Geographical load balancing with renewables," in *Proc. of the ACM GreenMetrics Workshop*, San Jose, CA, Apr. 2011.

[15] https://www2.ameren.com/RetailEnergy/realtimeprices.aspx.

[16] Http://ita.ee.lbl.gov/html/contrib/WorldCup.html.