Profit Maximization and Power Management of Green Data Centers Supporting Multiple SLAs

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

***Abstract*— While a large body of work has recently focused on reducing data center's energy expenses, there exists no prior work on investigating the *trade-off* between minimizing data center's energy expenditure and maximizing their revenue for various Internet and cloud computing services that they may offer. In this paper, we seek to tackle this shortcoming by proposing a systematic approach to maximize green data center's profit, i.e., revenue minus cost. In this regard, we explicitly take into account practical *service-level agreements* (SLAs) that currently exist between data centers and their customers. Our model also incorporates various other factors such as availability of local renewable power generation at data centers and the stochastic nature of data centers' workload. Furthermore, we propose an optimization-based profit maximization strategy for data centers for both cases, without and with behind-the-meter renewable generators. Using various experimental data and via computer simulations, we assess the accuracy of the proposed mathematical model for profit and also the performance of the proposed optimization-based profit maximization strategy.**

***Keywords*: Green data centers, service-level agreements, energy and performance management, renewable power generation.**

I. Introduction

Due to the increasing cost of electricity associated with data centers, there has been a growing interest towards developing techniques and algorithms to minimize data centers' energy expenditure. One thread of research focuses on reducing the amount of energy consumed by computer servers [1]. Another thread of research is *dynamic cluster server configuration* to reduce the total power consumption by consolidating workload only on a subset of servers and turning off the rest, during low workload hours [2], [3]. A similar approach is *dynamic CPU clock frequency scaling*, where a higher frequency, imposing higher energy consumption, is chosen only at peak workload hours [4]. Finally, some recent studies aimed to utilize *price-diversity* in deregulated electricity markets and *locational-diversity* in renewable power generation. They constantly monitor the price of electricity and the amount of renewable power generated at different regions and forward the workload towards data centers that are located in regions with the lowest electricity price [5] or highest renewable power available [6].

In this paper, we address the *trade-off* between *minimizing data center's energy expenditure* and *maximizing their revenue* for various Internet and cloud computing services that they may offer. Such trade-off is due to the fact that minimizing data center's energy cost is achieved essentially by turning off certain number of computers, scaling down CPU clocks, or migrating some workload, which can all potentially lead to degrading the quality-of-services offered by data center and consequently its income. In this regard, we propose a systematic approach to *maximize green data center's profit*, i.e., revenue minus cost. We explicitly take into account *service-level agreements* (SLAs) between data centers and their customers. Our model supports the common scenario that each data center may simultaneously offer different types of SLAs with different quality-of-service requirements. Our contributions in this paper can be summarized as follows:

- We develop a mathematical model to capture the trade-off between minimizing a data center's energy cost versus maximizing the revenue it receives for offering various Internet services. In this regard, we take into account computer server's power consumption profiles, data center's power usage effectiveness, price of electricity, availability of renewable power generation, behind-the-meter renewable generation contract models, total workload for each service class in terms of the rate at which service requests received at each time of day, different service-level agreements and their parameters including service deadline, service payment, and service violation penalty.
- We propose an *optimization-based profit maximization* strategy for data centers for both cases, without and with behind-the-meter renewable generators. The latter is the scenario applicable to green data centers.
- We use various *experimental and practical data*, e.g., for workload, price of electricity, renewable power generation, and SLA parameters, to assess the accuracy of the proposed mathematical model for profit and also the performance of the proposed optimization-based profit maximization strategy via computer simulations.

The rest of this paper is organized as follows. The system model and notations are defined in Section II. The proposed optimization-based profit maximization strategy is presented in Section III. Simulation results are presented in Section IV. Conclusions and future work are discussed in Section V.

II. System Model

A. Power Consumption

Consider a data center with M_{max} servers. Let $M \leq M_{max}$ denote the number of servers that are switched 'on'. 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 total electric power consumption associated with the data center can be obtained as [7]:

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

where U is the CPU utilization of servers and PUE is the *power usage effectiveness* of data center. 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 and Renewable Power Generation

To support both regulated or deregulated electricity markets, in our system model, we assume that the instantaneous price of electricity is denoted by ω. In Section III, we will use pricing information to obtain data center's cost of electricity.

To reduce cost of electricity, a data center may be equipped with behind-the-meter renewable generators, e.g., a wind turbine, in addition to being connected to the power grid. Let G denote the renewable power generated by renewable generators. The amount of power exchange with the grid is obtained as $P - G$. If local renewable power generation is lower than local power consumption, i.e., $P > G$, then $P - G$ is positive and the power flow is in the direction from the grid to the data center. While we allow a data center to inject its excessive renewable generation into the power grid, we assume that it does not receive compensation for the injected power.

C. Service Classes

A data center may offer different classes of service, e.g. cloud-based, video streaming, web services, etc. Each class has its specific method of service and SLA. We assume that a total of $N \geq 1$ service classes are offered by the data center.

D. Quality-of-Service

All arriving service requests are first received by a front-end web server, then forwarded and placed in one of the N depending on their class of service. The service requests of each class are then served in a first-come first-served order. To satisfy quality-of-service, the waiting time/queuing delay for each incoming service request of class i should be limited to a deadline that is determined by its corresponding SLA. Each SLA is identified by three parameters D, δ, and γ [7]. For the ith SLA, parameter D_i indicates the maximum waiting time (i.e., the deadline) that a service request of class i can tolerate. Parameter δ_i indicates the service money that the data center receives when it handles a single service request of class i before deadline D_i. Finally, parameter γ_i indicates the penalty that the data center has to pay to its customers every time it *cannot* handle a service request of class i before deadline D_i.

E. Service Rates

Let μ_i denote the rate at which service requests of class i are removed from their queue to be handled by a server. The total number of switched on servers is obtained as [7]

$$M = \sum_{i=1}^{N} \frac{\mu_i}{\kappa_i}. \quad (2)$$

As we increase the number of switched on servers and accordingly the service rates, more service requests can be handled before the SLA-deadline D_i, which in turn increases the payments that the data center receives as explained in Section II-D. On the other hand, it will also increase the data center's power consumption and accordingly the data center's energy expenditure as explained in Sections II-A and II-B. Therefore, there is a *trade-off* when it comes to selecting the data center's service rates, as we will discuss in detail next.

III. Problem Formulation

In general, the rate at which a data center receives *incoming service requests* of each service class can vary from time to time. To improve data center's performance, the number of switched on servers M should be adjusted according to the rates of receiving service requests, along with some other factors such as the hourly price of electricity and the amount of behind-the-meter renewable power generated. However, because of the tear-and-wear cost of switching on/off computers, M should not be changed too frequently. It is rather desired to be updated only every few minutes. Therefore, we divide running time of data center into several time slots $\Lambda_1, \Lambda_2, \cdots, \Lambda_N$ of length T, e.g. $T = 15$ minutes. The number of switched on servers are updated only at the beginning of each time slot. For the rest of this section, we focus on mathematically modeling the energy cost and profit of the data center of interest at each time slot $\Lambda \in \Lambda_1, \Lambda_2, \cdots, \Lambda_N$ as a function of service rates μ_i and consequently as a function of M, based on (2).

A. Revenue Modeling

Let $q_i(\mu_i)$ denote the probability that the queue waiting time for an incoming service request of class i exceeds the ith SLA-deadline D_i. Obtaining an analytical model for $q_i(\mu_i)$ requires a queueing theoretic analysis that we will provide in detail later in Section III-C. Let λ_i denote the average rate of receiving service requests of class i at the data center within the time slot Λ. The total revenue collected by the data center at the time slot of interest can be calculated as

$$Revenue = \sum_{i=1}^{N} (1 - q_i(\mu_i))\, \delta_i \lambda_i T - q_i(\mu_i)\gamma_i \lambda_i T, \quad (3)$$

where the first term within the summation, i.e., $(1 - q_i(\mu))\delta_i \lambda_i T$ denotes the total payments received by the data center within time interval T for the service requests of class i that are handled before the ith SLA-deadline, while the second term, i.e., $q_i(\mu)\gamma_i \lambda_i T$ denotes the total penalty paid by the data center within time interval T for the service requests of class i that are not handled before the ith SLA-deadline.

B. Cost Modeling

Within time interval T, each turned on server handle

$$\frac{T(1 - q_i(\mu_i))\lambda_i}{M} \quad (4)$$

service requests of class i. This makes each server busy for

$$\frac{T(1 - q_i(\mu_i))\lambda_i}{\kappa_i M} \quad (5)$$

seconds handling service requests of class i within the time interval T. Therefore, by dividing the total CPU busy time

$$\sum_{i=1}^{N} \frac{T(1-q_i(\mu_i))\lambda_i}{\kappa_i M} \tag{6}$$

by T, the CPU utilization for each server is obtained as

$$U = \sum_{i=1}^{N} \frac{(1-q_i(\mu_i))\lambda_i}{\kappa_i M}. \tag{7}$$

Replacing (2) and (7) in (1), the power consumption associated with the data center at the time slot of interest is obtained as

$$P = \sum_{i=1}^{N} \left[\frac{a\mu_i + b\lambda_i(1-q_i(\mu_i))}{\kappa_i} \right], \tag{8}$$

where

$$a \overset{\triangle}{=} P_{idle} + (E_{\text{usage}} - 1)P_{peak}, \tag{9}$$

$$b \overset{\triangle}{=} P_{peak} - P_{idle}. \tag{10}$$

Multiplying (8) by the electricity price ω, the total energy cost at the time interval of interest is obtained as

$$Cost = T\omega \sum_{i=1}^{N} \left[\frac{a\mu_i + b\lambda_i(1-q_i(\mu_i))}{\kappa_i} \right]. \tag{11}$$

C. Probability Model of $q_i(\mu_i)$

We model the ith SLA-deadline by a *finite size queue* with the length $\mu_i D_i$ [7], [8]. A service request of class i can be handled before the ith SLA-deadline, if and only if it enters the aforementioned ith finite size queue. We assume that the ith service request rate has mean λ_i and an arbitrary *general* probability distribution. On the other hand, since the service rate μ_i is fixed over each time interval of length T, $q_i(\mu_i)$ is modeled as the *loss probability* of a G/D/1 queue. Therefore, following the queuing theoretic analysis in [9], we can obtain

$$q_i(\mu_i) = \alpha_i(\mu_i)\, e^{-\frac{1}{2} \min\limits_{n \geq 1}\; m_{n,i}(\mu_i)}, \tag{12}$$

where

$$\alpha_i(\mu_i) = \frac{1}{\lambda\sqrt{2\pi}\sigma_i} e^{\frac{(\mu-\lambda_i)^2}{2\sigma_i^2}} \int_{\mu_i}^{\infty} (r-\mu_i) e^{-\frac{(r-\lambda_i)^2}{2\sigma_i^2}}\, dr \tag{13}$$

and for each $n \geq 1$ we have

$$m_{n,i}(\mu_i) = \frac{(D_i\mu_i + n(\mu_i - \lambda_i))^2}{nC_{\lambda_i}(0) + 2\sum\limits_{l=1}^{n-1} C_{\lambda_i}(l)(n-l)}. \tag{14}$$

Here, $\sigma_i = C_{\lambda_i}(0)$ and $C_{\lambda_i}(l)$ denotes the auto-covariance with lag time l of the rate of the incoming service request of ith class. The model in (12) has its most accuracy when the ith service request arrival rate is as a Gaussian process. However, it also works well for any general probability distribution [9], as we will confirm through simulations in Section IV.

D. Profit Maximization without Local Renewable Generation

At each time slot Λ, data center's profit is obtained as

$$Profit = Revenue - Cost, \tag{15}$$

where revenue is as in (3) and cost is as in (11). We seek to choose the data center's service rates μ_i to maximize profit. This can be expressed as the following optimization problem:

$$\begin{aligned} \underset{\lambda_i \leq \mu_i}{\textbf{Maximize}}\quad & T\sum_{i=1}^{N} [1-q_i(\mu_i))\delta_i\lambda_i - q_i(\mu_i)\gamma_i\lambda_i] - \\ & T\omega \sum_{i=1}^{N} \left[\frac{a\mu_i + b\lambda_i(1-q_i(\mu_i))}{\kappa_i} \right] \end{aligned} \tag{16a}$$

Subject To

$$\sum_{i=1}^{N} \frac{\mu_i}{\kappa_i} \leq M_{max}, \tag{16b}$$

where $q_i(\mu_i)$ is as in (12) and $M_{\max}$ denotes the maximum number of servers that are available in the data center. We note that, for each service class i, the service rate μ_i is lower bounded by λ_i. This is necessary to assure stabilizing the service request queues [7], [9]. We also note that the N-variable optimization problem in (16) can be decomposed into N separate optimization problems whenever $M_{\max}$ is too large. Finally, we note that Problem (16) needs to be solved once for every time slot $\Lambda \in \{\Lambda_1, \ldots, \Lambda_N\}$, i.e., every T minutes using a simple exhaustive search.

E. Profit Maximization with Local Renewable Generation

When a data center is supplied by both power grid and a local behind-the-meter renewable generator, then the optimum choices of service rates for maximizing profit is obtained by solving the following optimization problem

$$\begin{aligned} \underset{\lambda_i \leq \mu_i}{\textbf{Maximize}}\quad & T\sum_{i=1}^{N} [1-q_i(\mu_i))\delta_i\lambda_i - q_i(\mu_i)\gamma_i\lambda_i] - \\ & T\omega \left[\sum_{i=1}^{N} \frac{a\mu_i + b\lambda_i(1-q_i(\mu_i))}{\kappa_i} - G \right]^{+} \end{aligned} \tag{17a}$$

Subject To

$$\sum_{i=1}^{N} \frac{\mu_i}{\kappa_i} \leq M_{max}, \tag{17b}$$

where $[x]^+ = \max(x, 0)$. Note that, in Problem (17), optimization variables are coupled not only in constraint (17b), but also in objective function (17a) due to the term $[x]^+$.

IV. Performance Evaluation

A. Simulation Setting

Consider a data center and assume that the number of switched on servers M is updated periodically at the beginning of each time slot of length $T = 15$ minutes by solving Problems (16) and (17), for the cases without and with behind-the-meter renewable power generation at the data center,

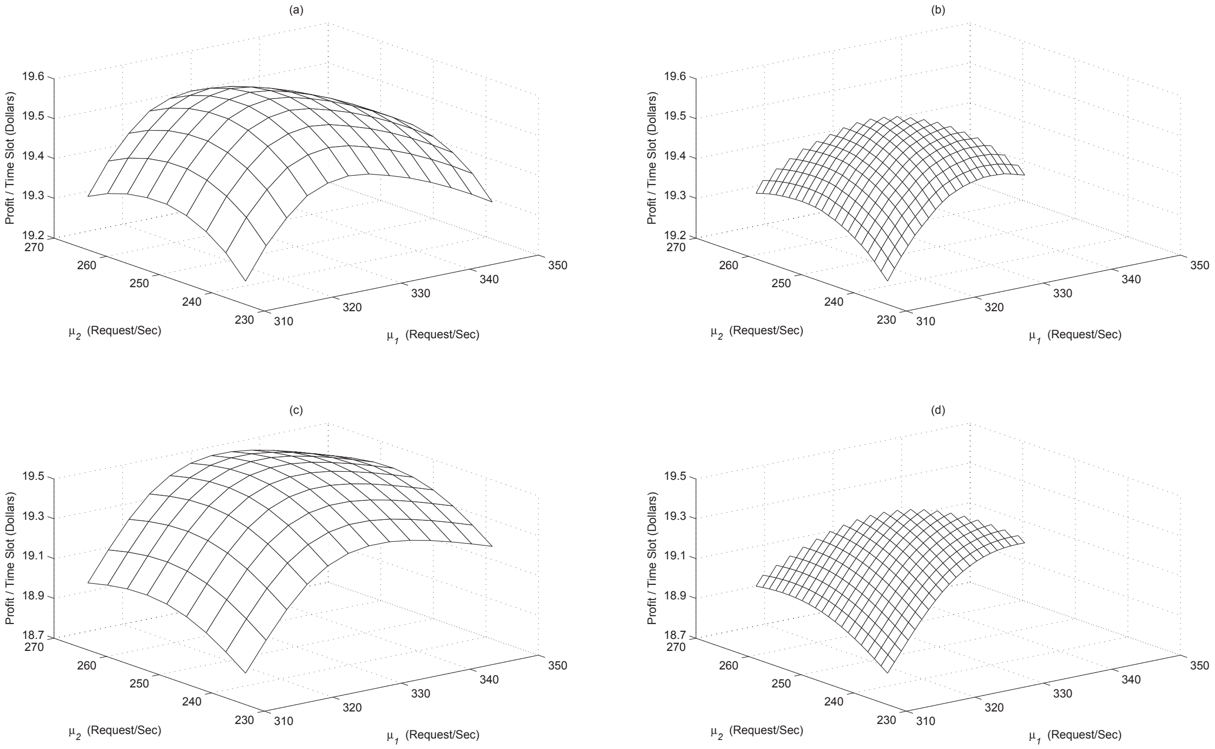

Fig. 1. Comparing analytical and simulation results to calculate data center profit. (a) Simulation results when parameter $M_{max} = 15000$. (b) Simulation results when parameter $M_{max} = 5250$. (c) Analytical results when parameter $M_{max} = 15000$. (d) Analytical results when parameter $M_{max} = 5250$.

respectively. For each switched on server, we have P_{peak} = 200 watts and P_{idle} = 100 watts [6]. We assume that $E_{\text{usage}} = 1.2$, which is the reported state of the art power usage effectiveness [7]. The electricity price information in our simulations are based on the real-time pricing tariffs currently practiced in Illinois Zone I, on June, 19, 2011 [10]. The prices are updated every hour. We assume that the data center offers $N = 2$ classes of services, which are identical to the Gold and Silver services explained in the example SLA model in [7]. We also assume that $\kappa_1 = .1$ and $\kappa_2 = .125$. To simulate the total workload, we use the publicly available World Cup 98 web hits data on June 14, 1998 and June, 19, 1998 as the service request trend for service class 1 and service class 2, respectively [11].

B. Simulation Results for a Single Time Slot

To gain insights about the achievable profit, in this section, we focus on a single time slot of length $T = 15$ minutes starting at 2:00 PM and we investigate the solution of the profit maximization problem in (16). The simulation versus analytical results for different choices of system parameter M_{max} are shown in Fig. 1. For the results in Fig. 1(a) and (c), we have $M_{\max} = 15000$. We can see that, since the number of available servers is large, constraint (16b) is *not* binding and the choices of optimization variables μ_1 and μ_2 can almost arbitrarily increase. However, given the trade-off between minimizing energy expenditure and maximizing revenue, optimal solution enforces switching on only a limited number of servers. Furthermore, we can see that the simulation curve in Fig. 1(a) and the analytical curve in Fig. 1(c) are very similar, suggesting that our proposed analytical profit model is accurate. For the results in Fig. 1(b) and (d), we have $M_{\max} = 5250$. We can see that, since the number of available servers is small, constraint (16b) *is* binding. As a result, the optimal solution is achieved when $\mu_1/\kappa_1 + \mu_2/\kappa_2 = M_{\max}$. That is, at optimality, we need to switch on all servers that are available in the data center. Nevertheless, we can see that although the feasible set for optimization problem (16) is significantly smaller when $M_{\max} = 5250$, the profit curve remains concave, again indicating the trade-off between minimizing energy expenditure and maximizing revenue.

C. Simulation Results for an Entire Day

For the rest of this section we assume that $M_{max} = 15000$. Simulation results over 24 hours are shown in Fig. 2. The time-of-day price of electricity is as in Fig. 2(a) [12]. We can see that the normalized profit gain in Fig. 2(b) is very close to optimal with 97% optimality on average. The normalized profit gain is calculated as $(Profit - Profit_{\text{Base}})$ divided by $(Profit_{\text{Max}} - Profit_{\text{Base}})$. Here, $Profit_{\text{Base}}$ is the profit obtained when we simply set $\mu_i = \lambda_i$ [13] and $Profit_{\text{Max}}$ is the maximum of the profit curve obtained by simulation.

D. Impact of Behind-the-Meter Renewable Generation

Next, assume that the data center is equipped with a wind turbines, with 1.5 Megawatts peak generation. The power

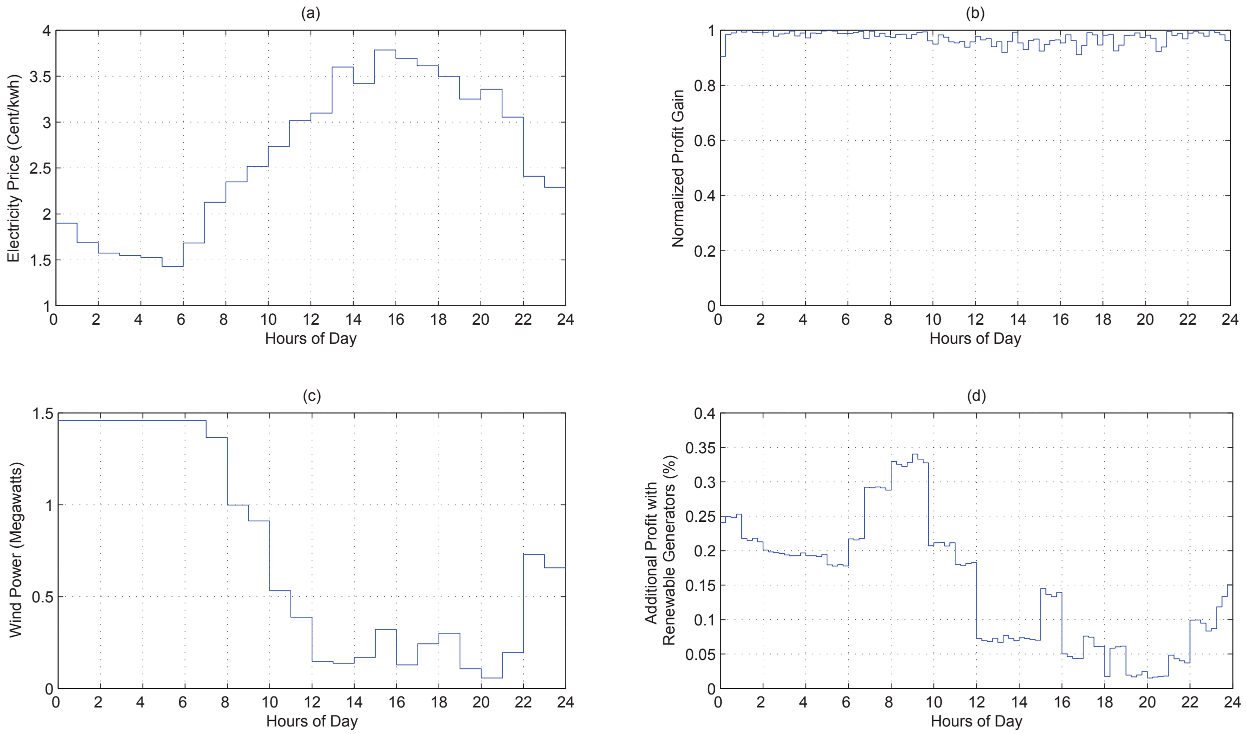

Fig. 2. Experimental data and the simulation results for 24 hours operation of the data center: (a) Time-of-Day Prices [12]. (b) Obtained normalized profit gain without local renewable power generation. (c) Available wind power [14]. (d) Additional profit (in percentage) with local renewable power generation.

output for each turbine is assumed to be as in Fig. 2(c), based on the wind speed data available in [14] for June 14, 2011. In this case, optimal service rate is obtained by solving Problem (17). The corresponding additional profit gain (in percentage) due to local renewable generation is shown in Fig. 2(d). We can see that local renewable generation can significantly increase the data center's daily profit.

V. Conclusions and Future Work

In this paper, we proposed a novel analytical model to calculate profit in large data centers without and with behind-the-meter renewable power generation. Our proposed model takes into account several factors including the practical service-level agreements that currently exist between data centers and their customers, price of electricity, and the amount of renewable power available. We then used the derived profit model to develop an optimization-based profit maximization strategy for data centers. Using various experimental data and via computer simulations, we assess the accuracy of the proposed mathematical model for profit and also the performance of the proposed optimization-based profit maximization strategy.

The results in this paper can be extended in several directions. First, the cost model can be extended to include cost elements other than energy cost. Second, the proposed energy and performance management method can be further extended to daily or monthly planning, e.g., to decide whether installing more servers is economical. Finally, the system model can be adjusted to also include potential profit if a data center participates in ancillary services in the power grid.

References

[1] S. Steinke, N. Grunwald, L. Wehmeyer, R. Banakar, M. Balakrishnan, and P. Marwedel, "Reducing energy consumption by dynamic copying of instructions onto onchip memory," in *Proc. of the International Symposium on System Synthesis*, Kyoto, Japan, Oct. 2002.

[2] J. Heo, D. Henriksson, X. Liu, and T. Abdelzaher, "Integrating adaptive components: An emerging challenge in performance-adaptive systems and a server farm case-study," in *Proc. of the IEEE International Real-Time Systems Symposium*, Tucson, AZ, Dec. 2007.

[3] 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.

[4] X. Fan, W. D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *Proc. of the ACM International Symposium on Computer Architecture*, San Diego, CA, June 2007.

[5] 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 IEEE INFOCOM*, Orlando, FL, 2010.

[6] 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.

[7] ——, "Energy and performance management of green data centers: A profit maximization approach," *submitted to IEEE Trans. on Smart Grid*, May 2012.

[8] ——, "Data centers to offer ancillary services," in *Proc. of the IEEE SmartGridComm12*, Tainan City, Taiwan, 2012.

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

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

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

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

[13] X. Zheng and Y. Cai, "Reducing electricity and network cost for online service providers in geographically located internet data centers," in *Green Computing and Communications (GreenCom), 2011 IEEE/ACM International Conference on*, aug. 2011, pp. 166 –169.

[14] http://www.windenergy.org/datasites.