Coordination of Cloud Computing and Smart Power Grids

Amir-Hamed Mohsenian-Rad and Alberto Leon-Garcia
Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada
e-mails: {h.mohsenian.rad, alberto.leongarcia}@utoronto.ca

***Abstract*—The emergence of cloud computing has established a trend towards building massive, energy-hungry, and geographically distributed data centers. Due to their enormous energy consumption, data centers are expected to have major impact on the electric grid by significantly increasing the load at locations where they are built. However, data centers and cloud computing also provide opportunities to help the grid with respect to robustness and load balancing. To gain insights into these opportunities, we formulate the *service request routing* problem in cloud computing jointly with the *power flow analysis* in smart grid and explain how these problems can be related. Simulation results based on the standard setting in the IEEE 24-bus Reliability Test System show that a grid-aware service request routing design in cloud computing can significantly help in *load balancing* in the electric grid and making the grid more reliable and more robust with respect to link breakage and load demand variations.**

I. Introduction and Motivation

Cloud computing has been envisioned as the next-generation computing paradigm for its major advantages in on-demand self-service, ubiquitous network access, location independent resource pooling, and transference of risk [1]. The main element in cloud computing is a shift in the geography of computation from the network edges to the Internet, i.e., the *cloud*. The *cloud providers* own large data centers with massive computation and storage capacities. They sell these capacities *on-demand* to the *cloud users* who can be software, service, or content providers for the users over the web [2].

The major cloud providers such as Google, Microsoft, and Amazon have built and are working on building the world's largest data centers across the United States and elsewhere. Each data center includes hundreds of thousands of computer servers, cooling equipment, and substation power transformers. For example, consider Microsoft's data center in Quincy, Washington. It has 43,600 square meters of space and uses 4.8 kilometers of chiller piping, 965 kilometers of electric wire, 92,900 square meters of drywall, and 1.5 metric tons of backup batteries. The company does not release the number of servers at this site; however it says that the data center consumes 48 megawatts which is enough to power 40,000 homes [3]. As another example, the National Security Agency is planning to build a massive data center at Fort Williams in Utah which is expected to consume over 70 megawatts electricity [4].

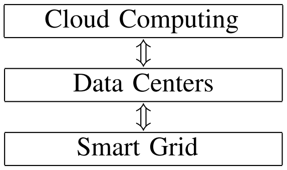

Fig. 1. The interactions between cloud computing systems and smart grid through massive, energy-hungry, and geographically distributed data centers.

The interactions between cloud computing, data centers, and smart grid are illustrated in Fig. 1. On the one hand, data centers are key elements of the cloud computing [2]. They are also expected to have major impact on the Internet and affect routing and congestion control algorithms [5]. On the other hand, due to their enormous energy consumption, data centers are expected to have major impact on the electric grid by increasing the load at locations they are built. Moreover, as data centers grow in size, the cost of electricity is dominating all other cost aspects in cloud computing. This leads to an increasing interest in devising resource management algorithms among data centers that take into account power grid-related issues such as the changes in electricity price during the day at different regions with different time-zones by dynamically *shifting* the computation load towards data centers which are located in regions with cheaper electricity [6], [7].

In this paper, unlike most of the previous work which are concerned with the negative impact of data centers' extra load to the electric grid, we would like to answer this question: is it possible to design cloud computing algorithms that can actually *help* smart grid design in terms of *load balancing* and *robustness*? We consider the *service request routing* algorithms that determine the distribution of the computation load among data centers. Since the computation load on each data center directly affects the data center's energy consumption [8], [9], we argue that we can reroute service requests towards different data centers in order to *control* cloud computing's impact on the power grid at different locations. For example, we can reduce the computation load and consequently the energy consumption at a data center in an area where the grid is prone to *circuite overflow*. In this regard, we formulate the service request routing problem in cloud computing jointly with the *power flow analysis* in smart grid within an *optimization framework* and explain how these problems are related. Our simulation results based on the settings in the IEEE 24-bus Reliability Test System show that a grid-aware service request routing design in cloud computing can significantly help for better load balancing and more robustness in the electric grid.

The rest of this paper is organized as follows. The system model and notations are described in Section II. Our proposed optimal grid-aware service request routing design is developed in Section III. Simulation results are presented in Section IV. Conclusions and future work are discussed in Section V.

II. System Model

In this section, we introduce the notation and system models for both power as well as data networks. We will use these models to formulate a grid-aware design optimization problem for cloud computing systems in Section III.

A. Power Network

Consider a power grid and let $\mathcal{N}$ with size $N = |\mathcal{N}|$ denote the set of all *buses*. The buses are interconnected through *branches* forming the grid topology. Each bus $i \in \mathcal{N}$ may also be connected to one or more generators or various loads. In our system model, some loads to the power grid may include large data centers which support cloud computing. Focusing on the *per-unit* setting of power distribution networks, we can derive *DC-equivalent power flow equations*[1] as follows [10]:

$$P_i = \sum_{j=1, j \neq i} B_{ij} \left(\theta_i - \theta_j \right), \qquad \forall\, i \in \mathcal{N}, \tag{1}$$

Here, for each bus $i \in \mathcal{N}$, parameter P_i denotes the amount of *active power injection* (i.e., total generation minus total load) to the grid at bus i, B_{ij} denotes the imaginary term in the complex value at row i and column j of the *Y-bus* matrix of the grid, and θ_i denotes the angle of the voltage phaser at bus i. In power flow equations, the only variables are angles θ_i for all buses $i \in \mathcal{N}$. In practice, one bus is selected as *slack bus* with zero phaser angle. Therefore, the phaser angles at all other buses are selected in terms of their differences with respect to the reference phaser angle in the slack bus [10].

Given the phaser angles $\theta_1, \ldots \theta_N$ obtained by solving the *system of linear equations* in (1), we can calculate the *active power flow* over each *branch* (i, j) of the power grid as

$$P_{ij} = B_{ij} \left(\theta_i - \theta_j \right). \tag{2}$$

The amount of P_{ij} directly affects the problem of circuit overflow in a power grid. That is, overflow occurs if the active power at branch (i, j) reaches its maximum permitted level $P^{\max}$. Therefore, it is required to always limit P_{ij} below the level $P^{\max}$. We should clarify that when it comes to branches among buses in a power distribution network, e.g., branch (i, j), *reactive power* Q_{ij} is significantly less than active power P_{ij}. Therefore, in practice, circuite overflow problem only involves branch active powers [10]. In summary, whether or not circuite overflow occurs in a power grid depends on the grid topology, the Y-bus matrix, and the amount of active power injection or consumption at all buses in the system:

$$\boldsymbol{P} \triangleq \begin{bmatrix} P_1 \\ \vdots \\ P_N \end{bmatrix}. \tag{3}$$

We note that the power injection/consumption at each bus may *change over time* due to i) changes in power generation capacity of the power plants especially those which use renewable energy sources, ii) changes in residential, commercial, and industrial load, and iii) changes in power consumption at large data centers connected to the grid. The last item is expected to grow over the next few years. In general, we have

$$\boldsymbol{P} = \boldsymbol{P}^{\text{Background}} + \boldsymbol{P}^{\text{DataCenter}}, \tag{4}$$

where

$$\boldsymbol{P}^{\text{Background}} \triangleq \begin{bmatrix} P_1^{\text{Background}} \\ \vdots \\ P_N^{\text{Background}} \end{bmatrix}, \tag{5}$$

and

$$\boldsymbol{P}^{\text{DataCenters}} \triangleq \begin{bmatrix} P_1^{\text{DataCenters}} \\ \vdots \\ P_N^{\text{DataCenters}} \end{bmatrix}. \tag{6}$$

We note that for each electric bus $i \in \mathcal{N}$, the term $P_i^{DataCenter}$ denotes the power consumption at the data centers (if any) connected to bus i; and the term $P_i^{Background}$ denotes any load *other than* data centers at bus i. Next, we model the data network and obtain expressions for data centers energy consumption as functions of their computation load.

B. Data Network

Consider several data centers connected to the Internet to support cloud computing. Each data center is also connected to one *bus* in the power grid to obtain the electricity needed for its operation. Let $\mathcal{S} \subseteq \mathcal{N}$ denote the set of buses in the grid that feed at least one data center. Also let $\mathcal{U} \subseteq \mathcal{N}$ denote the set of *user locations*. They represent cities and towns. We assume there is a bus at each user location to provide electricity for users as well. For each user location $u \in \mathcal{U}$ and each data center location $s \in \mathcal{S}$, we denote λ_{us} as the total *service requests* at user location u *routed* towards data center bus s. Service requests may range from simple query hits to web servers to extensive computation services for research purposes. Let L_u denote the total number of service requests at each user location u. We assume that the data centers are *fully replicated* [6], [7]. Therefore, to assure responding to all service requests from all users, it is required that we have

$$\sum_{s \in \mathcal{S}} \lambda_{us} = L_u, \qquad \forall\, u \in \mathcal{U}. \tag{7}$$

That is, all service requests at each user location need to be routed towards some data center. For the data center at bus $s \in \mathcal{S}$, the total number of service requests is obtained as

$$\sum_{u \in \mathcal{U}} \lambda_{us}. \tag{8}$$

Let μ denote the total number of service requests that a computer server can handle. Also let m_s denote the number of servers at the data center connected to bus $s \in \mathcal{S}$. The corresponding *average server utilization* is defined as

$$\gamma_s \triangleq \left(\sum_{u \in \mathcal{U}} \lambda_{su} \right) / \left(m_s\, \mu \right), \qquad \forall\, s \in \mathcal{S}. \tag{9}$$

Clearly, server utilization always needs to be less than one to ensure handling all service requests in a timely manner.

[1] The currents in the grid are *not* direct; however due to the complexity of AC power flow equations, the more simplified DC-equivalent power flow equations are commonly used in power flow analysis in practice [10], [11].

Let P_{idle} denote the average *idle power* draw of a single computer server and P_{peak} denote the average *peak power* when the server is handling a service request. In addition, we denote *power usage effectiveness* (PUE)[2] by E_{usage} [12]. The ratio P_{peak}/P_{idle} denotes the *power elasticity* of the servers. A higher value of this ratio indicates greater elasticity, leading to less power consumption when the server is idle. We can obtain the total power consumption corresponding to the data center connected to each bus $s \in \mathcal{S}$ as [8]:

$$P_s^{\text{DataCenter}} \triangleq m_s \left(P_{idle} + (E_{usage} - 1) \times P_{peak} \right) + m_s \left(P_{peak} - P_{idle} \right) \times \gamma_s + \epsilon, \tag{10}$$

where ϵ is an empirical constant. The expression in (10) has two key terms. The first term, i.e., $m_s(P_{idle} + (E_{usage} - 1) \times P_{peak})$ represents the *base usage* which does *not* depend on the computation load. The second term, i.e., $m_s(P_{peak} - P_{idle}) \times \gamma_s$, represents the *added usage* which indicates the extra power consumption depending on the computation load.

The relationship between cloud computing's service request routing and smart grid's circuites overflow problems are now evident. In fact, we can change power consumption at each bus connected to a data center by changing the distribution of the service requests among data centers. This will directly impact the power load on different branches in the grid following the power flow equations. Therefore, we can *control* cloud computing's impact on the electric grid at different locations by changing service request routing distribution in the system:

$$\boldsymbol{\lambda} \triangleq \left(\lambda_{su}, \; \forall \, s \in \mathcal{S}, u \in \mathcal{U} \right). \tag{11}$$

Finally, for notational simplicity, we assume that

$$P_i^{\text{DataCenter}} = 0, \qquad \forall \, i \notin \mathcal{S}. \tag{12}$$

That is, if an electric bus i is not feeding any data center, then the corresponding data center load at bus i is simply zero.

III. Grid-Aware Service Request Routing in Cloud Computing for Power Load Balancing

Given the system model and the relationships between service request routing algorithms and the circuite overflow problem that we explained in Section II, we are now ready to introduce our design to have a cloud computing system which can *help* load balancing and robustness in the electric grid.

Recall that for each branch (i,j) in the grid the power transmission load P_{ij} is obtained as in (2). Clearly, the higher the value of P_{ij}, particularly the closer $P_{i,j}$ is to $P_{ij}^{\max}$, the electric branch (i,j) would be more prone to circuite overflow in case of background load variations or power link breakage. Therefore, it is usually desired to reduce the power transmission on branches and have them significantly less than their allowed capacity. That is, for each branch (i,j), we would like to keep the following fraction as low as possible:

$$P_{ij}/P_{ij}^{\max}. \tag{13}$$

In this regard, a reasonable *load balancing* design objective can be formulated in terms of solving the following optimization problem across *all* the branches in the electric grid:

$$\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\textbf{minimize}} & \max\limits_{(i,j)} \; P_{ij}/P_{ij}^{\max} \\ \textbf{subject to} & \text{Eqs. } (1) - (12). \end{array} \tag{14}$$

The optimization variables in problem (14) are the portion of service requests to be routed towards each data center in the cloud computing system. In contrast, the design objective is entirely a smart grid-related concept. To better understand how the service request routing strategies would affect load balancing and the design objective in problem (14), let us look at the following *directional relationship* expressions:

$$\boldsymbol{\lambda} \;\Rightarrow\; \boldsymbol{P}^{\text{DataCenters}} \;\Rightarrow\; \boldsymbol{P} \;\Rightarrow\; \max_{(i,j)} \; P_{ij}/P_{ij}^{\max}. \tag{15}$$

From (4), (9), (10), and (11) it is evident that any changes in the service request routing vector variable $\boldsymbol{\lambda}$ can potentially lead to changes in vector $\boldsymbol{P}^{\text{DataCenters}}$ and consequently changes vector $\boldsymbol{P}$. On the other hand, from the power flow equations (1) and (2), changes in bus powers $\boldsymbol{P}$ will also change branch powers at different locations in the grid and can lead to different values for $\max_{(i,j)} \; P_{ij}/P_{ij}^{\max}$. Due to the enormous energy consumption at data centers these changes can be major and significantly affect load balancing in the electric grid.

We note that the optimization problem in (14) can be transformed easily to a standard *linear programming problem* (cf. [13]). To see this, let $\Gamma \geq 0$ denote an *auxiliary variable* added to the system. We can rewrite optimization problem (14) as the following *equivalent* optimization problem [14]:

$$\begin{array}{lll} \underset{\boldsymbol{\lambda},\, \Gamma \geq 0}{\textbf{minimize}} & \Gamma & \\ \textbf{subject to} & P_{ij}/P_{ij}^{\max} \leq \Gamma, & \forall \, i,j \in \mathcal{N}, \\ & \text{Eqs. } (1) - (12). & \end{array} \tag{16}$$

It is easy to verify that the objective function and all constraints in optimization problem (16) are *affine*. Thus, problem (16) is indeed a linear programming problem which can be solved efficiently using techniques such as the *simplex method* or the *interior point method* [13]. We note that the computational complexity of solving the linear programming problem (16) is *not* a concern as data centers have very advanced computation capabilities and they can easily update the service request routing plans on a frequent basis, whenever they obtain new information about the background load in the grid.

As a special case, assume that all branches have the same maximum capacity. That is, $P_{ij}^{\max} = P^{\max}$ for all $i,j \in \mathcal{N}$. In that case, the load balancing problem (14) becomes

$$\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\textbf{minimize}} & \max\limits_{(i,j)} \; P_{ij} \\ \textbf{subject to} & \text{Eqs. } (1) - (12). \end{array} \tag{17}$$

The equivalent linear program for problem (17) can be obtained accordingly. Intuitively, the optimal solutions of the load balancing problems (14)-(17) would move the computation load to data centers in the areas of the grid where generation capacity is relatively higher than the background load. This will avoid the need for major power transmission across long distances and through multiple branches in the grid.

[2] A measure of data center energy efficiency. Currently, the typical value for most enterprise data centers is 2.0 or more. However, the studies have suggested that by the year of 2011 most data centers could reach a PUE of 1.7. A few state-of-the art facilities could reach a PUE of 1.2 [12].

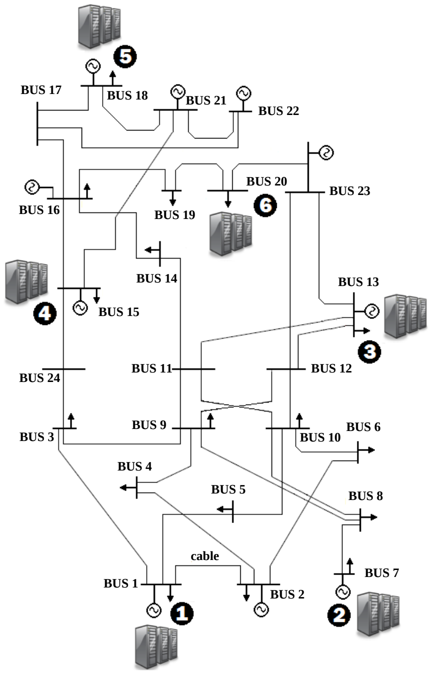

Fig. 2. The IEEE 24-bus reliability test system added with six data centers. For each data center, energy consumption depends on computation load.

IV. Simulation Results

In this section, we assess the performance of the optimal solutions obtained by solving the load balancing optimization problem (14). We show that the proposed grid-aware service request routing algorithm for cloud computing systems can lead to significantly more balanced power load distribution among electric branches and a more robust smart grid design.

A. Simulation Setting

Consider the power grid in Fig. 2. This is a slightly modified version of the IEEE 24-bus reliability test system introduced in [15]. In total, there are 24 buses and 38 branches in the system. There are 10 buses with generation capacity and 17 buses with background load demand. Unless we state otherwise, the generation capacities and background loads are assumed to be as shown in Table I. These values resemble the data provided at the IEEE 24-bus reliability test system standard.

There are also six data centers connected to the grid at buses 1, 7, 13, 15, 18, and 20. Each data center is equipped with 600 thousands computer servers and can have up to 100 megawatts load at its peak energy consumption level. We have P_{peak} = 140 watts and P_{idle} = 40 watts [8]. We also have $E_{usage} = 1.2$ which is the reported state of the art power usage effectiveness [12]. Each server can handle one request per minute, which implies that $\mu = 60$. For the purpose of our study and without lack of generality, we assume that in total the data centers are expected to handle 100 million service requests per hour. This number is reasonable compared to the query hit rates at large cloud providers such as Google [16]. Recall that depending on the distribution of how service requests are routed towards different data centers, the energy consumption level at each data center may change. Finally, for the purpose of our study, we assume the use of underground monopole *high voltage direct current* (HVDC) transmission lines, where for each branch (i, j) we have $P_{ij}^{\max}$ = 600 megawatts [17].

TABLE I
Generation Capacity and Background Load at Each Bus[‡]

	Generation Capacity	Background Load
BUS 1	172	8[†]
BUS 2	172	100
BUS 3	-	180
BUS 4	-	74
BUS 5	-	71
BUS 6	-	136
BUS 7	115	25[†]
BUS 8	-	171
BUS 9	-	175
BUS 10	-	195
BUS 11	-	-
BUS 12	-	-
BUS 13	286	165[†]
BUS 14	-	194
BUS 15	215	217[†]
BUS 16	155	100
BUS 17	-	-
BUS 18	250	233[†]
BUS 19	-	181
BUS 20	-	28[†]
BUS 21	348	-
BUS 22	300	-
BUS 23	660	-
BUS 24	-	-

‡ All amounts are in megawatts.
† To be added by load of a data center.

B. Impact of Grid-aware Cloud Computing

Based on the choices of parameters described in Section IV-A, the *base usage* at each data center is 40 megawatts and the total *added usage* to be distributed among all six data centers is 180 megawatts. If the computation load is *evenly* distributed among the data centers then the energy consumption at each data center becomes 40 + 180 / 6 = 70 megawatts. In that case, the highest power transmission load on a branch would be 330 megawatts which occurs on branch $(21, 15)$ on the direction from bus 21 to bus 15. However, if the service requests are routed according to the *solution* of problem (14), then *no* service request would be routed towards the first, second, and fourth data centers. In that case, the highest power transmission load on a branch would reduce to 300 megawatts which again occurs on branch $(21, 15)$. We can see that the proposed design can reduce the transmission load on the *bottleneck branch* by about 10% leading to a noticeably better load balancing across the power grid. This is because the load from cloud computing is *shifted* from the lower part of the grid in Fig. 2, where load demand is higher than the generation capacity, to the upper part of the grid, where the generation capacity is higher than the demand load. We note that the performance improvement can potentially be higher in larger networks when cloud computing load is more significant.

C. Daily Trend

Next, let us look at the results when the power network in

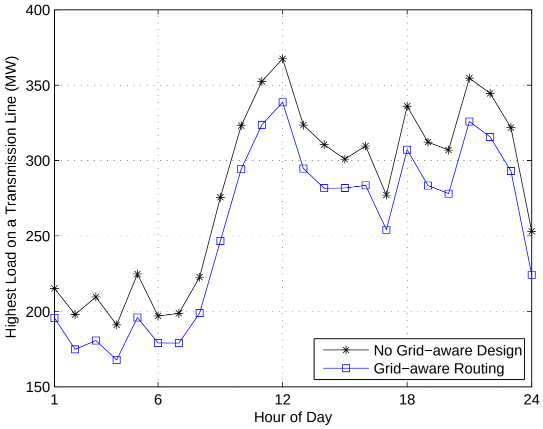

Fig. 3. The daily trend of the transmission load of the bottleneck electric branch with and without the use of the proposed grid-aware design.

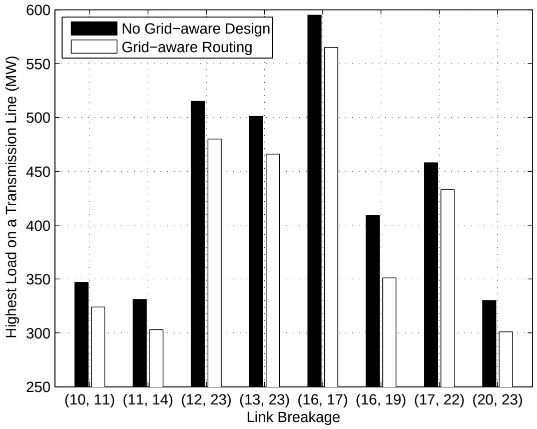

Fig. 4. The impact of link breakage on transmission load of the bottleneck electric branch with and without the use of the proposed grid-aware design.

Fig. 2 operates for the whole day. We assume that the daily load profile is based on the data provided in the IEEE 24-bus reliability test system standard [15] plus some randomness in background load at each bus. Simulation results in terms of the highest transmission load among the branches are shown in Fig. 3. We can see that for different background load scenarios and at different hours of the day, the proposed design can help for better load balancing across the branches in the power grid.

D. Link Breakage

Finally, we investigate the impact of *link breakage*. We examine the scenarios where exactly *one* electric branch fails and goes out of service and then we look at the resulted transmission load on all the *other* branches across the grid. Some examples are shown in Fig. 4. We can see that in most cases a link breakage can drastically increase the highest load among branches compared to the results we saw in normal operation in Fig. 3. This can put the electric grid at the risk of circuit overflow. For example, when link breakage occurs at branch $(16, 17)$, the transmission power on the bottleneck branch $(21, 15)$ increases up to 595 megawatts which is very close to the 600 megawatts maximum permitted load on this branch. Nevertheless, we can see that in all the considered link breakage scenarios, the proposed grid-aware cloud computing design can help to reduce the transmission load spikes on the rest of branches in the studied electric grid. This leads to a more robust and more reliable smart grid design.

V. Conclusions

This paper represents one of the first steps towards understanding the interactions between cloud computing and smart grid through the algorithms which involve massive data centers. We focused on one design possibility that can improve load balancing in the grid by carefully distributing the service requests among data centers in a clouding computing system. We took a systematic approach and formulated the service request routing problem in cloud computing jointly with the power flow analysis in the smart grid and explained how this can lead to grid-aware cloud computing routing algorithms. Simulation results based on the setting in the IEEE 24-bus Reliability Test System show that our design can significantly improve robustness and load balancing in a smart grid.

References

[1] B. Hayes, “Cloud Computing,” *Communications of the ACM*, vol. 51, no. 7, pp. 9–11, Jul. 2008.

[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, “Above the clouds: A berkeley view of cloud computing,” University of California at Berkeley, Research Report, Feb. 2009.

[3] R. H. Katz, “Tech Titans Building Boom,” *IEEE Spectrum*, pp. 40–54, Feb. 2009.

[4] R. Miller, “NSA Plans 1.6 Billion Dollars Utah Data Center,” Data Center Knowledge Website, Jun. 2009.

[5] H. Shimonishi, J. Higuchi, T. Yoshikawa, and A. Iwata, “A congestion control algorithm for data center area communications,” in *Proc. of IEEE International Communications Quality and Reliability Workshop*, Vancouver, Canada, Jun. 2010.

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

[7] 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*, San Diego, CA, Mar. 2010.

[8] X. Fan, W. D. Weber, and L. A. Barroso, “Power provisioning for a warehouse-sized computer,” in *ACM International Symposium on Computer Architecture*, San Diego, CA, Jun. 2007.

[9] A. H. Mohsenian-Rad and A. Leon-Garcia, “Energy-information transmission tradeoff in green cloud computing,” in *Proc. of the IEEE Globecom’10*, Miami, FL, Dec. 2010.

[10] A. J. Wood and B. F. Wollenberg, *Power Generation, Operation, and Control*. Wiley-Interscience, 1996.

[11] M. Aigner and E. Oswald, *Power Analysis Tutorial*. Austria: Institute for Applied Information Processing and Communication - University of Technology Graz, 1989.

[12] U.S. Environmental Protection Agency, *Server and Data Center Energy Efficiency - Final Report to Congress*, 2007.

[13] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. Belmont, MA: Athena Science, 1997.

[14] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Sci., 2004.

[15] “Reliability Test System Task Force of the Application of Probability Methods subcommittee - The IEEE Reliability Test System - 1996,” pp. 1010–1020, Aug. 1999.

[16] A. Lipsman, “Google Gets 76 Billion Searches Per Month of 113 Billion Total,” WebSite 100 Marketing Communications, Available at http://website101.com/press/google-76-billion-searches, Sep. 2009.

[17] R. Rudervall, J. P. Charpentier, and R. Sharma, “High Voltage Direct Current (HVDC) Transmission Systems Technology Review Paper,” World Bank, Mar. 2000.