Energy-Information Transmission Tradeoff in Green Cloud Computing

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*— With the rise of Internet-scale systems and cloud computing services, there is an increasing 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 and potentially the amount of greenhouse gas emissions and carbon footprint. In this regard, the locations that are selected to build future data centers as well as the service load to be routed to each data center after it is built need to be carefully studied given various environmental, cost, and quality-of-service considerations. To gain insights into these problems, we develop an optimization-based framework, where the objective functions range from minimizing the energy cost to minimizing the carbon footprint subject to essential quality-of-service constraints. We show that in multiple scenarios, these objectives can be conflicting leading to an energy-information tradeoff in green cloud computing.**

***Index Terms*— Green cloud computing, data centers, routing algorithms, electric grid, carbon footprint, price of electricity, price of bandwidth, carbon tax, renewable power, optimization.**

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 and most advanced data centers across the Unites States and elsewhere. Each data center includes hundreds of thousands of computer servers, cooling equipments, and one or more 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 batteries for backup power. 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].

One of the key questions that a cloud provider needs to answer is: *At which locations its data centers should be built?* Traditionally, it is argued that since the cost of wide-area networking is higher than all other IT hardware costs, economic necessity mandates putting the data near the users to minimize the cost of deploying long data links with high bandwidth [4]. However, as the data centers grow in size and become major energy consumers, the cost of electricity is now dominating all other costs including the cost of bandwidth [2]. As a result, most future data centers are expected to be built in locations where the electricity is inexpensive. For 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 in its two phases [5]. There is also an increasing interest in devising routing algorithms that take into account the changes in electricity prices during the day at different regions with different time-zones to dynamically shift the high computation load towards data centers which are located in regions with cheaper electricity [6]–[8].

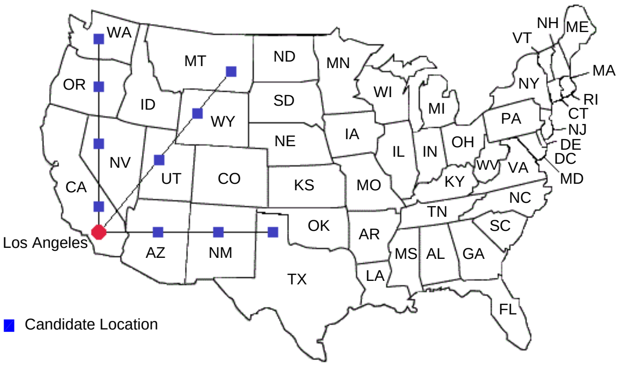

Fig. 1. The problem of selecting the best location to build a large data center for Los Angeles among ten candidate locations in ten different states.

While there is a growing attention to the cost of electricity associated with cloud computing, the environmental impacts of energy-hungry data centers are less addressed. In fact, cheap electricity can sometimes be at the cost of harm to the environment. Currently, *six* out of *ten* states with the lowest electricity prices in the United States have significantly higher *carbon footprint*[1] associated with their power sectors compared to the nation's average amount. For example, while Utah and Wyoming are ranked as the second and the third states with the lowest electricity prices [10], they are also ranked as the fourth and third states in terms of the normalized amount of carbon dioxide emissions per each kilowatt hour (kWh) electricity generation at their power plants [11], [12].

[1] Carbon footprint denotes the amount of the poisoning carbon dioxide (CO_2) gas emissions during the normal operation of electricity generation and consumption systems. It is usually higher for coal-fired and natural gas generators and lower for nuclear and hydroelectric power plants [9].

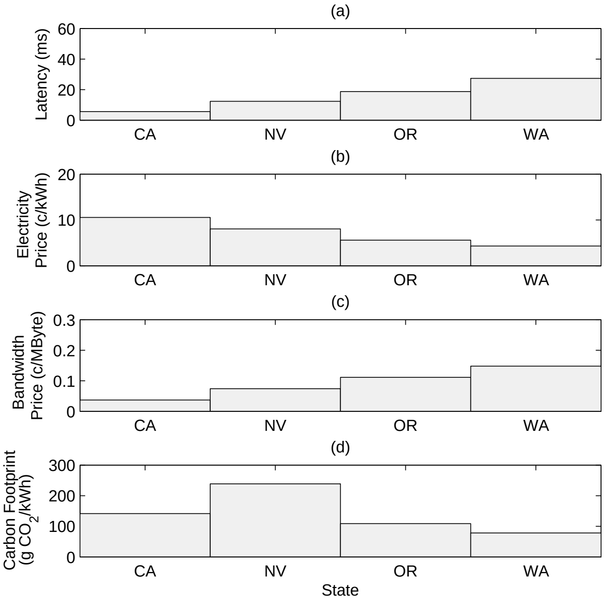

Fig. 2. Different factors that can help us compare the four candidate data center locations along the line from California to Washington in Fig. 1.

To better understand the trade-offs among cost, environmental, and quality-of-service design aspects in green cloud computing, consider the problem of selecting the best location to build a large data center for Los Angeles. This problem is illustrated in Fig. 1, where *ten* candidate locations are being examined. Considering the four candidate locations along the line from California to Washington, the corresponding latency, electricity price, bandwidth price, and carbon footprint are shown in Fig. 2, based on various data collected from [2], [10]–[12]. We can see that as we move away from Los Angeles, both latency and bandwidth price increase. Instead, we can see significant decrease in the price of electricity especially in Washington. We can also see that carbon footprint corresponding to the operation of the data center may vary independent of the price of electricity. For example, while moving the data center from California to Nevada will save 24% on energy cost it will also increase carbon footprint associated with the operation of this data center by 68%.

In this paper, we systematically study the energy-information transmission tradeoff in selecting the best locations to build data centers for green cloud computing. Our work is based on formulating various optimization problems, where the objective functions range from minimizing the energy and bandwidth cost to minimizing the total carbon footprint subject to quality-of-service constraints. The solutions of the formulated optimization problems determine: 1) whether or not we should build a data center at each candidate location, 2) how many computer servers we would need to deploy in each data center, and 3) how the service requests from users in different places need to be routed towards each data center.

Our work in this paper is closely related to various studies in the literature. One thread of existing work addresses green data center design [13]. The focus is on efficiently utilizing the servers and improving the power management schemes used in data centers to reduce carbon footprint. Another thread of research focuses on energy-cost aware request routing among data centers by considering the geographically dependent electricity cost [6], [7]. While we also consider the cost of electricity, our design here is more general as we decide on not only the request routing, but also the data center locations and the number of servers in each data center. Moreover, we consider the cost associated with *carbon tax* which is not taken into account in any prior work in the literature. Finally, some design aspects that we discuss here, such as building a data center close to renewable energy sources, have been mentioned in various panel discussions and magazine articles, e.g., in [14]. However, here we address these environmental aspects within a more systematic and analytical framework.

Paper Organization: We introduce the system model in Section II. Three optimization problems to select data center locations are formulated in Section III. Simulation results are given in Section VI. The paper is concluded in Section V.

II. System Model

In this section, we introduce the system model which we will later use to formulate various design optimization problems to select data center locations in section III.

A. Candidate and User Locations

Consider a single cloud provider and let $\mathcal{S}$, with size $S = |\mathcal{S}|$, denote the set of all *candidate locations* for building data centers. For each candidate location $s \in \mathcal{S}$, we define $x_s = 1$ if a data center is actually built in location s. Otherwise, $x_s = 0$. Assuming that the total number of data centers that are planned to be built is $X \ll S$, then it is required that we have

$$\sum_{s \in \mathcal{S}} x_s = X. \tag{1}$$

On the other hand, let $\mathcal{U}$, with size $U = |\mathcal{U}|$, denote the set of all *user locations*. They represent cities and towns. We may also aggregate the users in close-by regions into a single representative user location. For each user location $u \in \mathcal{U}$ and for each hour of the day $h \in \mathcal{H} \triangleq [1, \ldots, 24]$, the total expected number of web service requests per hour is denoted by L_u^h. Note that the number of requests can drastically change during the day. For example the measurements by Akamai Technologies show that its content distribution servers receive around four times more hits per second at high-peak hours in the evening compared to off-peak hours in early morning [6].

B. Service Request Routing

Let λ_{su}^h denote the portion of load L_u^h which is planned to be routed towards data center location $s \in \mathcal{S}$ at each hour $h \in \mathcal{H}$. To assure responding to all requests, it is required that[2]

$$\sum_{s \in \mathcal{S}} \lambda_{su}^h = L_u^h, \qquad \forall\, h \in \mathcal{H}. \tag{2}$$

For each $s \in \mathcal{S}$, any $u \in \mathcal{U}$, and each $h \in \mathcal{H}$, we further define $y_{su}^h = 1$ if the data center which is built in candidate location s is planned to handle *any* request from user location

[2] Here we assume that the data centers are *fully replicated* [6], [7].

u at hour h. Otherwise, $y_{su}^h = 0$. Therefore, we need to have

$$0 \leq \lambda_{su}^h \leq y_{su}^h \, L_u^h, \qquad \forall \, s \in \mathcal{S}, \; u \in \mathcal{U}, \; h \in \mathcal{H}, \tag{3}$$

and

$$y_{su}^h \leq x_s, \qquad \forall \, s \in \mathcal{S}, \; u \in \mathcal{U}, \; h \in \mathcal{H}. \tag{4}$$

From (3), if no service request from user location u is routed at a certain hour h to the potential data center in candidate location s, then we would automatically have $\lambda_{su}^h = 0$. On the other hand, from (4), if no data center is built in candidate location s, then no request from any user location can be routed to this candidate location resulting in $\lambda_{su}^h = 0$ for *all* user locations $u \in \mathcal{U}$ at *all* hours of the day $h \in \mathcal{H}$.

C. Number of Computer Servers

Let m_s denote the number of computer servers in a candidate location $s \in \mathcal{S}$. We assume that if candidate location s is selected to host a data center, m_s should be lower and upper bounded by $M^{\min}$ and $M^{\max}$, respectively. That is,

$$x_s \, M^{\min} \leq m_s \leq x_s \, M^{\max}, \qquad \forall \, s \in \mathcal{S}. \tag{5}$$

From (5), if $x_s = 0$, then $0 \leq m_s \leq 0$, i.e., $m_s = 0$. That is, if no data center is planned to be built in candidate location s, essentially no computer server will be deployed in that location either. On the other hand, if $x_s = 1$, then the corresponding constraint in (5) simply becomes $M^{\min} \leq m_s \leq M^{\max}$.

D. Power Consumption and Availability

Let P_{idle} denote the average *idle power* draw of a single 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)[3] by E_{usage} [15]. The ratio P_{peak}/P_{idle} denotes the *power elasticity* of the servers. The higher the value of this ratio indicates more elasticity, leading to less power consumption when the server is idle. We can obtain the total power consumption at each candidate location $s \in \mathcal{S}$ and at each hour of the day $h \in \mathcal{H}$ as [16]:

$$\begin{aligned} P_s^h = \; & m_s \left(P_{idle} + (E_{usage} - 1) \times P_{peak} \right) \\ & + m_s \left(P_{peak} - P_{idle} \right) \times \gamma_s^h + x_s \, \epsilon, \end{aligned} \tag{6}$$

where ϵ is an empirically derived constant and γ_s^h denotes the *average server utilization* at hour h which is obtained as

$$\gamma_s^h = \left(\sum_{u \in \mathcal{U}} \lambda_{su}^h \right) / \left(m_s \, \mu \right). \tag{7}$$

Here, μ denotes the total number of service requests that a computer server can handle in one hour. We note that if $P_{peak} = P_{idle}$, then $P_s^h = m_s \; E_{usage} \; P_{peak} + \epsilon$ and power consumption would depend solely on the number of servers, not the number of routed requests or the hour of the operation. We also note that if $x_s = 0$ and no data center is actually built in candidate location n, then $P_s^h = 0$ for all $h \in \mathcal{H}$.

Depending on the number of power plants in a region and their capacities as well as the existing residential, commercial, and industrial load, there is a limited power available to run data centers at each candidate location. As an example, the total annual electricity generation capacity in California is 54 million megawatt-hours *less* than the total annual load in this state. In contrast, total annual electricity generation capacity in Wyoming is 30 million megawatt-hours *more* than the total annual load in this state [12]. Therefore, Wyoming has more power available for large data centers. This can be taken into account by introducing maximum available power $P_s^{h,\max}$ at each user location and at each hour $h \in \mathcal{H}$ such that

$$P_s^h \leq P_s^{h,\max}, \qquad \forall \, s \in \mathcal{S}, \; h \in \mathcal{H}. \tag{8}$$

Note that the power available at a candidate location may vary during the day due to two reasons. First, some power plants, e.g., those with renewable energy sources such as wind and sun light, have time-varying generation capacity. Second, most power plants are not used exclusively by data centers and they need to provide electricity for various other load subscribers as well. However, the electricity consumption by both residential and commercial users is time-varying which leads to changes in available power to run data centers at different hours.

E. Cost of Electricity and Bandwidth

The electric grid in North America is operated on a regional basis. In a few regions with *deregulated* electricity market, prices may differ at different hours of the day as they reflect the wholesale electricity market in the region. The prices are usually higher at peak hours in the evenings and lower at night. On the other hand, in most areas electricity market is *regulated* and prices are fixed and do not change during the day. In our system model, we consider the general case where in each candidate location $s \in \mathcal{S}$ and at each hour $h \in \mathcal{H}$, the price of electricity is denoted by θ_s^h. As a result, the corresponding hourly cost of electricity can be obtained as $\theta_s^h \, P_s^h$.

The cost of bandwidth depends on the distance between user location and candidate location and the amount of data to be transmitted. In our model, we define σ_{su} as the hourly-equivalent cost of transmitting a unit of bandwidth between user location u and candidate location s. This represents the total cost of bandwidth amortized over several years.

Besides the cost of electricity and bandwidth, other major costs for data centers are the costs of cooling and property and equipment purchase [2]. The cost of equipment is independent of the location of the data center. One can also find inexpensive land outside cities and next to rivers to reduce the property purchase and cooling costs almost in all states. From this and since our focus is on energy and environmental aspects of cloud computing, we do not take these costs into consideration.

F. Power Generation and Carbon Footprint

Let $\rho_s \in [0, 1]$ denote the power transmission *loss rate* at each candidate location $s \in \mathcal{S}$. In general, loss rate is higher for longer transmission lines when the load is located far from major power plants in the region. The loss rate also depends on the electric grid topology. Given the loss rates, we can

[3] 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 [15].

obtain the *total power generation* needed to run a data center at candidate location s and at each hour $h \in \mathcal{H}$ as $(1+\rho_s)\, P_s^h$.

Depending on the location of the load, the required electricity can come from different types of power plants with different emission levels. Recall that carbon footprint is higher for coal-fired and natural-gas generators and lower for nuclear and hydroelectric generators [9]. By considering the major power plants around each candidate location $s \in \mathcal{S}$, we can obtain the average carbon dioxide emission level ϕ_s corresponding to the generation of 1 kWh electricity in this candidate region. Clearly, if a data center is located in an area such as Wyoming or Utah where electricity is almost entirely generated by coal-fired power plants, then although electricity cost would be low, it is at the cost of harm to the environment.

G. Carbon Tax

In order to enforce environmental considerations, *carbon tax* is currently used in a few states in the United States, such as Colorado, and multiple provinces in Canada, such as Quebec and British Columbia [17], [18]. In most instances, carbon tax is applied to power plants, which they in turn pass the cost of the carbon price onto consumers by increasing the price of electricity. In that case, environmental enforcements would be taken into consideration in our system model within the cost of electricity. However, since carbon taxes are not popular yet, we introduce them as separate parameters in our studies in order to understand their impacts. In this regard, for each candidate location $s \in \mathcal{S}$, we denote the carbon tax by δ_s which leads to an *additional* payment of $\delta_s\,(1+\rho_s)\,P_s^h$. As an example, consider a 70 megawatts data center near the city of Boulder in Colorado where the highest carbon tax rate is 0.49 cents per kWh [18]. Even if we assume no loss, a daily operation of this data center under full utilization would lead to more than 8000 dollars for carbon tax each day, which is noticeable.

H. Quality-of-Service Requirements

Let D_{su} denote the propagation delay between a data center in candidate location $s \in \mathcal{S}$ and a user location $u \in \mathcal{S}$. In order to avoid long response delays in handling service requests, we assume that the *round trip* propagation delay is always required to be upper bounded by $D^{\max}$. That is, we have

$$2\, D_{su}\, y_{su}^h \leq D^{\max}, \qquad \forall\, s \in \mathcal{S},\ u \in \mathcal{U},\ h \in \mathcal{H}. \tag{9}$$

Note that if $y_{su}^h = 0$, i.e., no service request is routed from user location s to candidate location u at hour h, the corresponding inequality in (9) would hold regardless of the value for D_{su}.

n the other hand, to limit the queuing delay, i.e., the waiting time for a service request at data center before it is handled by a server, we limit *average server utilization* at each data center by a constant $\gamma^{\max} \in (0,1]$. That is, we have

$$\gamma_s^h \leq \gamma^{\max}, \qquad \forall\, s \in \mathcal{S},\ h \in \mathcal{H}. \tag{10}$$

The choice of parameter $\gamma^{\max}$ depends on service request traffic pattern and the quality-of-service requirements, e.g., see [7]. In general, if $\gamma^{\max}$ is small enough, then the waiting time at servers would be negligible and mostly the propagation delay, which is already bounded by $D^{\max}$, would form the overall latency in responding to service requests.

III. Optimal Data Center Location Selection and Service Request Routing

In this section, we formulate three different optimization problems in order to decide on the right location to build each data center. For notational simplicity, we define $\mathbf{x} \triangleq [x_s, \forall\, s \in \mathcal{S}]$, $\mathbf{y} \triangleq [y_{su}, \forall\, s \in \mathcal{S},\ u \in \mathcal{U}]$, $\mathbf{m} \triangleq [m_s, \forall\, s \in \mathcal{S}]$, and $\boldsymbol{\lambda} \triangleq [\lambda_{su} \forall\, s \in \mathcal{S},\ u \in \mathcal{U}]$. Thus, the optimization variables in our designs would be $\mathbf{x}$, $\mathbf{y}$, $\mathbf{m}$, and $\boldsymbol{\lambda}$. They determine whether or not we should build a data center in each candidate location, how many servers we need in each data center, and how the service requests would need to be routed towards data centers.

A. Design I: Minimizing Total Carbon Footprint

In order to minimize the total carbon dioxide emissions associated with the operation of all data centers, we would need to solve the following optimization problem:

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{y},\mathbf{m},\boldsymbol{\lambda}}{\textbf{minimize}} & \displaystyle\sum_{s\in\mathcal{S}}\sum_{h\in\mathcal{H}} \phi_s\,(1+\rho_s)\,P_s^h \\ \textbf{subject to} & \text{Eqs. } (1)-(8). \end{array} \tag{11}$$

We note that minimizing the objective function in problem (11) can be done by not only reducing power consumption in data centers but also by building them in areas with *clean* electricity generation such as in regions with large hydroelectric dams.

B. Design II: Minimizing Total Cost With Carbon Tax

While reducing the carbon footprint would benefit the society, it is natural to expect that what cloud providers are interested in is to reduce their operation cost not necessarily the environmental overhead. Nevertheless, we can still enforce environmental considerations by applying carbon tax which is directly related to the carbon footprint associated with operation of each data center. In this regard and in order to minimize the total daily cost of running the data centers, we would need to solve the following optimization problem:

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{y},\mathbf{m},\boldsymbol{\lambda}}{\textbf{minimize}} & \displaystyle\sum_{s\in\mathcal{S}}\sum_{h\in\mathcal{H}} \Bigg(\theta_s\,P_s^h + \delta_s\,(1+\rho_s)\,P_s^h \\ & \displaystyle\qquad\qquad + \sum_{u\in\mathcal{U}} \lambda_{su}^h\,\sigma_{su}\Bigg) \\ \textbf{subject to} & \text{Eqs. } (1)-(8). \end{array} \tag{12}$$

As we discussed in Section II-F, the major cost of carbon tax can potentially lead to reconsideration with respect to where data centers should be built and how service requests at various user locations should be routed towards different data centers.

C. Design III: Minimizing Average Service Latency

Considering the energy-information transmission trade-off, an alternative design objective can be minimizing the average latency in responding to service requests which is obtained as

$$\left(\sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}}\sum_{h\in\mathcal{H}} \lambda_{su}^h D_{su}\right) \Big/ \left(\sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}}\sum_{h\in\mathcal{H}} \lambda_{su}^h\right).$$

Since the denominator equals $\sum_{u \in \mathcal{U}} \sum_{h \in \mathcal{H}} L_u^h$ which is *constant*, we can formulate the problem of minimizing the average service latency as the following optimization problem:

$$\begin{aligned} \underset{\mathbf{x},\mathbf{y},\mathbf{m},\boldsymbol{\lambda}}{\textbf{minimize}} \quad & \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}} \sum_{h \in \mathcal{H}} \lambda_{su}^h D_{su} \\ \textbf{subject to} \quad & \text{Eqs. } (1) - (8). \end{aligned} \tag{13}$$

We can expect that the optimal solution of problem (13) encourages building the data centers closer to the users.

D. Solution Approaches

The optimization problems in (11)-(13) are *linear mixed integer programming* problems which can be solved using various optimization software such as CPLEX [19] and MOSEK [20]. They usually work based on variations of the *branch-and-bound* algorithm for integer programming [21]. There also exist multiple heuristic such as the *iterated local search* and other *metaheuristic* methods [22] which can be used to obtain efficient sub-optimal solutions by relaxing the integer constraints. Nonetheless, we believe that computational complexity is not a major concern in our design as deciding on the location of data centers can be done over offline computations.

IV. Simulation Results

A. Simulation Setting

Consider the contiguous United States with 48 states and District of Columbia. We first note that the states of California, Nevada, Idaho, South Dakota, Minnesota, Wisconsin, Ohio, Tennessee, Florida, North Carolina, Virginia, Maryland, New York, Delaware, New Jersey, Connecticut, Rhode Island, Vermont, Massachusetts, and District of Columbia have *shortage* in their local electricity generation compared to their local consumptions [12]. Thus, we exclude them from the list of candidate locations. On the other hand, considering the difference between total local generation and total consumption, we assume that available power for any data center in states of Colorado, Iowa, Mississippi, Kentucky, and Georgia is 60 megawatts while all the other 23 states can host large data centers with total consumption up to 100 megawatts, if needed.

The rest of the simulation parameters are selected as follows. We set $P_{peak} = 140$ watts, $P_{idle} = 84$ watts [16], and $E_{usage} = 2$ [15]. Each server can handle one request per second, which implies that $\mu = 3600$. We also set $\gamma^{\max} = 0.8$ [23]. The price of electricity is based on the average price for industrial load at each state as listed in [10]. As in [2], we assume that \$1 buys 2.7 GByte of bandwidth over 100 kilometers. The size of service request and response files are 10 KByte. Since carbon tax is not common in most states, we use the current 0.49 cents per kWh carbon tax in Colorado as a basis to obtain equivalent carbon taxes in all other states. For example, knowing that the average carbon dioxide emission for 1 kWh electricity generation in Wyoming is about 1.8 more than that in Colorado [11], the equivalent carbon tax in Wyoming is assumed to be $0.48 \times 1.8 = 0.88$ cent per 1 kWh consumption. For each candidate location and user location, the propagation delay is obtained proportional to the distance between the two locations. The propagation delay increases by 10 ms every thousand kilometers (at each direction from and to the data center) when fiber optic links are used. Without loss of generality, we assume that $D^{\max} = 70$ ms.

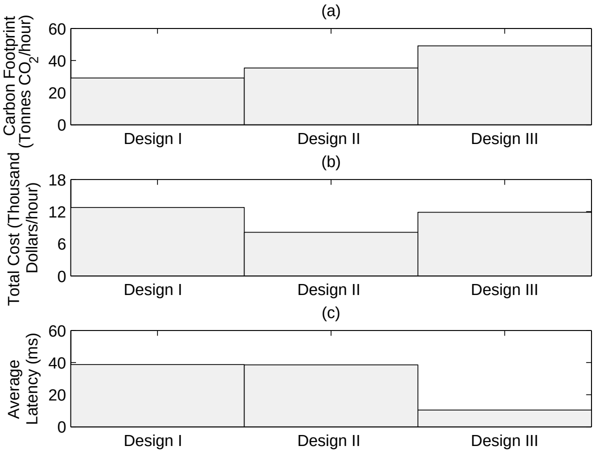

Fig. 3. Performance comparison among three design approaches introduced in Section III to select four data centers across United States.

For the purpose of our study, we assume that there are 100 different service types. The total number of hits for each service type is set according to the reported 293 million Google's domestic queries per day in [24]. For each state, the total number of daily service requests is adjusted to be proportional to the total state's population. Moreover, we assume that the distribution of the hits at each hour follows the daily trend of the number of hits to the content distribution servers reported by Akamai Technologies [6]. Finally, we take into account all four time zones in the contiguous United States to localize the daily hit rate distributions at each region.

B. Comparing Different Design Approaches

In this section, we compare Designs I, II, and III when they are used to find the best locations to build *four* data centers across the United States. The results are shown in Fig. 3 in terms of the carbon footprint, total cost which includes cost of electricity, cost of bandwidth, and carbon tax, and average latency. When Design I is used, the data centers would be built in Washington (100 megawatts), Mississippi (55 megawatts), New Hampshire (35 megawatts), and Oregon (35 megawatts). In this case, the total carbon footprint for all four data centers would become 29 tonnes per hour. The total cost in this case would be 12 thousand dollars per hour. On the other hand, when Design II is used, the data centers would be built in Washington (100 megawatts), Oklahoma (50 megawatts), Kentucky (20 megawatts), and Iowa (15 megawatts). The price of electricity is relatively low in all these states. Interestingly, we can see that due to carbon tax, states of Utah and Wyoming are *not* selected to build the data centers even though they are currently ranked as the second and the third states with the lowest electricity prices in the United States [10]. In this case, the carbon footprint would be 35 tonnes per hour which is 20% higher than the optimal case in Design I. Instead, the total cost can significantly reduce to 8 thousand dollars per hour. Finally, when Design III is used, the data centers would be built in Illinois (60 megawatts), Arizona (40 megawatts), Pennsylvania

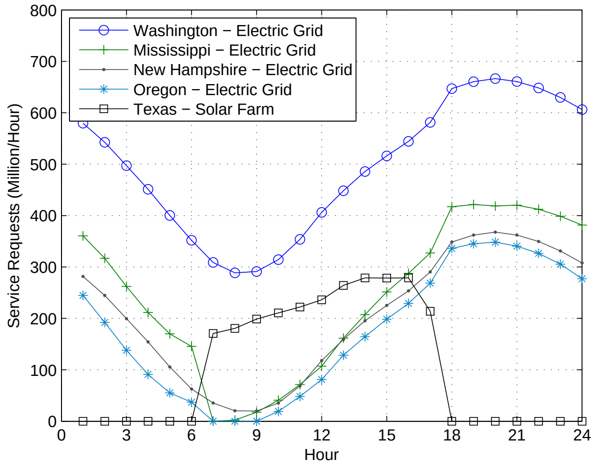

Fig. 4. Performance comparison among three design approaches introduced in Section III to select four data centers across United States.

(40 megawatts), and South Carolina (35 megawatts). In this case, the carbon footprint would be at the highest level among the three design approaches, i.e., 49 tonnes per hour; however, the average latency would become as low as only 10 ms.

C. Impact of Renewable Energy Generation

To obtain insights into the possibility of running data centers by renewable energy sources, we assume that besides the four data centers selected in Design I in Section IV-B, there is another data center next to a *solar farm* in Texas, which is *not* connected to the electric grid. The solar farm is assumed to be used exclusively by the data center and to have the peak capacity of 15 megawatts. The key challenge here is the fact that solar power is only available during the day, typically between 6:00 AM in the morning until 6:00 PM in the evening. The simulation results are shown in Fig. 4. In this figure, we plotted the total number of service requests routed towards each data center. We can see that the solar-powered data center can be used as long as it is available and the service requests would simply be routed to the other data centers when the solar power is not available. In this case, the carbon footprint reduces to 26 tonnes per hour which is 12% less than the case when all data centers are connected to the electric grid. These preliminary results suggest that it could be beneficial to build data centers close to renewable energy sources as cloud computing can easily handle variations in available power by *rerouting* the service requests towards different regions.

V. Conclusions and Future Work

In this paper, we systematically investigated the energy-information transmission tradeoff in green cloud computing. We provided an optimization-based framework, where the objective functions range from minimizing the energy and bandwidth cost to minimizing the total carbon footprint subject to quality-of-service constraints. We also provided initial results based on actual price and carbon footprint data in the United States and showed examples about how different design objectives can potentially lead to different and sometimes conflicting results in terms of deciding on where the data centers need to be built and how the service requests need to be forwarded towards different data centers. We also studied the impact of carbon tax on cloud computing and the possibility to run data centers by renewable power generation.

The results in this paper can be extended in various directions. First, it is interesting to extend the planning scope from *day* to *multiple years* to incorporate the impact of seasonal changes in electricity generation capacity and consumption as well as the predicted changes in various price aspects. Second, one can include the costs of cooling and property purchase and also the cost of renewable energy generation. Third, we can also take into account more elaborate quality-of-service requirements besides limiting the propagation delay.

References

[1] B. Hayes, “Cloud Computing,” *Communications of the ACM*, vol. 51, no. 7, pp. 9–11, July 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] J. Gray, “Distributed computing economics,” *Queue*, vol. 6, no. 3, pp. 63–68, May 2003.

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

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

[9] United Kingdom Parliamentary Office of Science and Technology, *Carbon Footprint of Elecricity Generation*, Oct. 2006.

[10] “Average Retail Price of Electricity to Ultimate Customers by End-Use Sector by State,” U.S. Energy Information Administration, Mar. 2010.

[11] “Carbon Dioxide Emissions From Power Plants Rated Worldwide,” Science Daily Website, Nov. 2007.

[12] “State Electricity Profiles - Summer Capacity,” U.S. Energy Information Administration, Mar. 2008.

[13] L. Liu, H. Wang, X. Liu, W. B. He, Q. B. Wang, and Y. Chen, “Green cloud: A new architecture for green data center,” in *Proc. of ICAC-INDST*, Barcelona, Spain, June 2009.

[14] R. Cocchiara, H. Davis, and D. Kinnaird, “Choosing data centre location,” *Datacenter Dynamics Knowledge Bank - Online*.

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

[16] 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, June 2007.

[17] “British Columbia Carbon Tax,” Ministry of Small Business and Revenue, Feb. 2008.

[18] “Climate Action Plan Tax,” City of Boulder - Colorado, Oct. 2009.

[19] “ILOG CPLEX.” http://www.ilog.com/products/cplex/, 2009.

[20] “MOSEK.” http://www.mosek.com, 2009.

[21] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Science, 2004.

[22] H. R. Lourenco, O. Martin, and T. Stutzle, “Iterated local search,” in *Handbook of Metaheuristics*, F. Glover and G. Kochenberger, Eds. Kluwer Academic Publishers, 2002, pp. 321–353.

[23] K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiron, F. Tobagi, and C. Diot, “Analysis of measured single-hop delay from an operational backbone network,” in *Proc. of IEEE INFOCOM*, New York, NY, June 2002.

[24] 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, Sept. 2009.