Balance your Bids before your Bits: The Economics of Geographic Load-Balancing

Jose Camacho
Universidad Carlos III de Madrid, Spain
jmc.camacho@gmail.com

Ying Zhang, Minghua Chen, Dah Ming Chiu
Dept. of Information Engineering
The Chinese University of Hong Kong
{zy013,minghua,dmchiu}@ie.cuhk.edu.hk

ABSTRACT

By routing workload to locations with cheaper electricity, geographic load-balancing (GLB) has been shown a promising mechanism to cut down the electricity bill of geo-distributed datacenters operated by the same organization. Most existing studies on GLB assume that the use of GLB has no impact on electricity prices, even though GLB increases local electricity demand variation. In practice, however, electricity prices are determined by how supply and demand are dynamically balanced by local electricity utilities, and thus may as well be affected by GLB. In this paper, in order to understand and unleash GLB's economic potential, we carry out a comprehensive study on how GLB interacts with electricity supply chains. In particular, we show that as GLB introduces extra uncertainty in local demand, utility companies may have to increase electricity prices to ensure certain profit margin in face of such demand uncertainty. Consequently, cloud service providers (CSP) doing GLB may end up getting minor cost reduction or even paying *higher* electricity bills than not doing GLB, as shown in our case study based on real-world traces. Then, motivated by the recent practice of large CSPs moving into electricity markets, we propose to allow CSPs to purchase electricity from markets through brokers. The advantage is that GLB no longer causes economic loss to utilities. Meanwhile, CSPs can still exploit their presence in multiple geo-locations to achieve desirable electricity cost reduction. Our case study using real-world traces shows that the solution can save CSPs up to 12% of the electricity cost.

Categories and Subject Descriptors

J.7 [**Computers in other systems**]: Industrial control ; G.1.6 [**Optimization**]: Stochastic Optimization

Keywords

smartgrid, datacenter, pricing, electricity market, auction, geographic load-balancing

e-Energy'14, June 11–13, 2014, Cambridge, UK.

http://dx.doi.org/10.1145/2602044.2602068 .

1. INTRODUCTION

The flourishing Internet-scale cloud services are revolutionizing the landscape of human activity. The rapid growth of such services has triggered an increasing deployment of massive geo-distributed data centers worldwide.

As a result, energy consumption of data centers hosting these services has been skyrocketing. In 2010, data centers worldwide consumed an estimated 240 billion kilowatt-hours (kWh) of electricity [18], almost enough to power the entire Spain [31]. The corresponding worldwide data center annual electricity bill is around 16 billion US dollars [18]. Today, energy cost represents a large fraction of the data center operating expense [9], and the cost is increasing at an alarming rate of 12% annually [34]. Consequently, reducing energy cost has become a critical concern for data center operators.

There have been a significant amount of academic and industrial efforts on minimizing data center energy cost; see for instance [28, 30, 27] and a recent survey in [6]. Among them, in this paper we focus on the solutions that exploit "price-aware" geographic load-balancing (GLB) across geo-distributed data centers.

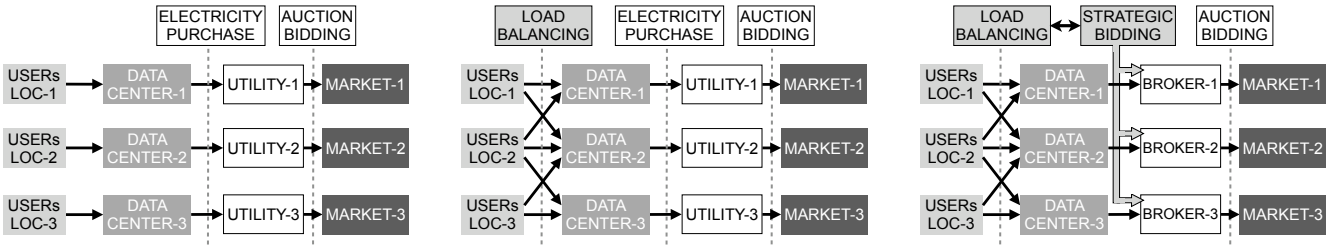
For cloud service providers (CSPs) that own data centers in different geographic locations, such as Google, Microsoft, and Amazon, routing user requests to locations with cheaper electricity has been shown a promising approach to cut down the electricity bill; see *e.g.*, [20, 36, 26, 33] and the references therein. These exciting studies suggest that GLB could achieve cost reduction (not necessary energy reduction) of 30-40%, depending on the flexibility of the service provider to shift traffic among locations.

Nevertheless, all existing works focus on addressing technical feasibility and revealing the abundant benefits of GLB, assuming the electricity prices are not affected by GLB, even though GLB increases local electricity demand variation.

In practice, however, the electricity prices are determined by how supply and demand are dynamically balanced by local utilities, and thus may as well be affected by GLB. In particular, the fact that the electricity is a non-storable commodity forces the utility to predict the demand and schedule its supply in advance. As GLB increases demand variation, it may incur extra errors in demand prediction. As we will show, these prediction errors will lead to over-/under-supply and consequently to economic loss of utilities. As a result, utilities may have to increase electricity prices to ensure certain profit margin in face of such extra economic loss caused by GLB.

Therefore, in order to understand and unleash GLB's economic potential, it is critical to understand the interaction between the GLB ability to alter electricity demand patterns, and the impact of this uncertainty on the electricity prices.

Before we turn to our focus and contributions, we note that *GLB can cause non-negligible demand variation for a utility*. For example, Facebook, Apple, Google and Amazon have built or will build



(a) Conventional electricity supply chain. (b) Electricity supply chain with GLB. (c) Geo-Distributed electricity supply chain.

Figure 1: Three electricity ecosystems studied in this paper.

large data centers in Prineville (Oregon, US) to leverage the chilly outdoor air for data center cooling at low cost. A fully-operated data center (*e.g.*, Google's data center in Oregon) is estimated to consume 90 MW power [5]. Power Pacific, a large utility serving Oregon including Prineville, sells 35 GWh daily [25]. Hence, these data centers once all in full operation could consume 8.6 GWh daily or 22% of Power Pacific sales today, and 33% in 4 years if we aggressively consider data center energy demand grows 15% annually as estimated in [18] while conventional demand remains steady. If data centers can shift 30% electricity demand away by doing GLB according to the estimate in [26], then GLB could lead to 10% demand variation for Power Pacific in 4 years.

Motivated by the above observations, we develop relevant models and carry out a comprehensive study of the impact of GLB on the electricity supply chain. Specifically, we analyze the intriguing interaction of GLB and utilities, revealing fundamental insights for the following two scenarios:

- **Current Model**: In this scenario (see Fig. 1(a)), electricity utilities purchase electricity from local electricity spot markets. Then, the utilities sell electricity like a commodity to data center owners to support their operation. The scenario evolves to Fig. 1(b) if GLB is used.

- **Broker Model**: In this scenario (see Fig. 1(c)), data center owners directly purchase electricity from local spot markets, either by obtaining a valid license[1] or through a broker (*e.g.*, utilities are ideal candidate for brokers).

In particular, we make the following contributions.

▷ We first give a brief overview of the electricity supply chain and introduce CSPs doing GLB as a *new* type of customers – they can make their local demand more *elastic* to prices by "shifting" electricity demand among geo-locations (Sec. 2). They are very different from conventional electricity customers whose demands are localized and inelastic.

▷ We provide a pricing model for the electricity sold by the utilities (Sec. 3). This model takes into account the increments in price to compensate the demand prediction errors and the price volatility from the market auctions.

▷ Then, motivated by the recent practice of large CSPs moving into electricity markets, we propose to allow CSPs to directly purchase electricity from markets through brokers (Sec. 4). By doing GLB and electricity procurement jointly, CSPs can eliminate the trading inefficiency between utilities and CSPs. Consequently, GLB no longer causes economic loss to utilities, and CSPs can still exploit their presence in multiple geo-locations to achieve desirable electricity cost reduction. Specifically, CSPs can first bid in different spot markets, *i.e.*, *balance their bids*, and then depending on their purchase of electricity perform GLB to optimize the load distribution, *i.e.*, *balance their bits*.

▷ In the performance evaluation section, by analysis and case study using real-world traces, we investigate the interaction of GLB with the supply chain and its economic consequence (Sec. 5). We show that electricity utilities rely on accurate demand prediction to balance supply and demand efficiently. As GLB makes accurate demand prediction harder, it causes trading inefficiency between utilities and CSPs and subsequently economic loss to the utilities. As a result, utilities will have to increase retail prices to ensure certain profit margin in face of the economic loss. Consequently, CSPs doing GLB end up getting poor cost reduction or even paying higher electricity bills than not doing GLB – 1% higher in our case study. The second part of the section shows that the broker-assisted GLB solution can save CSPs up to 12% of the electricity cost, while avoid the issues risen by the conventional (non-coordinated) GLB.

After discussing the related work in Sec.6, we conclude the paper in Sec.7.

2. THE ELECTRICITY SUPPLY CHAIN

In this section, we provide a high-level introduction of the electricity supply chain. In general, electricity supply chains consist of four components:

- *Generating Companies* (GENCOs),
- *Electricity Wholesale Market* (Market),
- *Utility Companies* (Utilities),
- *Customers* (in particular, Cloud Services Providers (CSPs) that owns multiple geo-distributed data centers).

Their interaction is shown in Fig. 1(a) (or Fig. 1(b) if CSPs perform GLB). GENCOs run the generating units and sell electricity on the wholesale Market. Utilities buy from the Market and sell retail to CSPs. For our study, it suffices to consider three components in the supply chain: Market, Utilities, and CSPs.

In the common practice today, the supply is traded in multiple timescales to match the demand. For example, in the US, the most common are day-ahead and real-time trading in the supply chain. Our study focuses on the day-ahead trading, which is based on a forward market that determines largely the hourly supply available to the utilities in the next day. The hourly timescale aligns with the suggested time granularity for CSPs to perform GLB [26].

[1]As a real-world example, in February 2010 the Federal Energy Regulatory Commission authorized Google to buy and sell energy at market rates [14].

2.1 Electricity Spot Markets

In recent years, the landscape of electricity wholesale trading has completely shifted towards de-regularized *spot markets*, to allow renewable energy integration and improve trading efficiency to offer lower prices to end customers [4].

In every spot market, the electricity supply is auctioned[2]. The *sellers*, *i.e.*, GENCOs, submit (hourly) generation offers, and the *buyers*, *i.e.*, Utilities, submit (hourly) demand bids, all in the form of <*marginal price, quantity*>, to the Independent System Operator (ISO), *i.e.*, the *auctioneer*. In the offers, the GENCOs specify the amount of electricity they want to sell (resp. Utilities specify in the bids the amount they want to buy) and at which marginal price. Each seller (resp. buyer) is allowed to submit multiple offers (resp. bids) in the same auction with different prices and quantities.

The ISO matches the offers with the bids, typically using a well-established double auction matching mechanism. The mechanism is rather sophisticated in details (we refer interested readers to [19, 16] and focus on the necessary background here), but the outcome is that it determines a *market clearing price* (MCP) for all the traded units.

The MCP clears the market in the following sense. A selling offer (<marginal price, quantity>) with the marginal price below the MCP is successful – the specified amount of electricity is sold on the market at the MCP. Thus *successful sellers sell at prices at least as good as what they offered*. Meanwhile, a buying bid succeeds if the buying price is above the MCP; then, the specified amount of electricity is purchased from the market at the MCP and *buyers pay no more than what they bid*. Remaining selling offers fail as their marginal prices are above the MCP (resp. remaining buying bids fail as their marginal prices are below the MCP[3]).

The MCP is jointly determined by independent bids submitted by uncoordinated parties. Because of the gigantic amount of electricity and capital involved in the auction, no single buyer or seller should dominate the market and determine the MCP. In practice, MCP can be well modeled as a random number drawn from an empirical distribution built from historical data, *independent of individual bids*. See later Fig. 2 for the empirical MCP distribution (ranging from 35 $/MWh to 130 $/MWh) of three day-ahead spot markets in the US.

2.2 Electricity Utilities

Similar to the retailers in a generic supply chain, utilities buy commodity – electricity – from spot markets and sell to CSPs to power data centers. Utilities make profit by selling electricity at a proper retail price. A conservative estimate of the retail prices for data centers today is about 60 $/MWh [26].

Meanwhile, utilities are unique retailers in two senses:

- utilities are trading a non-storable commodity (electricity) with very short "expiration time";
- utilities have to schedule electricity purchase one day before the demand arrives, by bidding in the day-ahead market.

These two facts force the utilities to *predict* precisely both the demand quantity and time-of-arrival, so as to *schedule* the purchase of the right amount of supply to be served at the right time. For example, a utility that predicts a data center needs 30MWh electricity tomorrow at 2-3pm needs to buy today, from the day-ahead market, the predicted amount of electricity for its dispatch tomorrow 2-3pm. If there are errors in the prediction, utilities will suffer from over-/under- supply. Over-/under- supply leads to either unmatched demand (to be compensated in more *volatile* markets) or unused electricity. Both immediately translates into economic loss for the utility.

Consequently, when setting the retail price, utilities have to take into account the potential economic loss due to demand prediction error. Larger demand uncertainty leads to larger prediction error, and thus higher economic loss. This observation is crucial in understanding the results in Sec. 3.

2.3 Cloud Services Providers (CSPs)

In this paper, we consider CSPs that operate energy-hungry geo-distributed data centers (*e.g.*, Google and Microsoft) to provide *computing-intensive* services (*e.g.*, search) to its users through the Internet. Depending on whether they perform GLB, CSPs' roles as electricity customers differ significantly.

- Without GLB, a CSP manages its geo-distributed data centers separately as shown in Fig. 1(a). Each data center only serves its regional workload, and it purchases electricity from local utilities for its energy needs. In this case, from the utilities' point of view, each data center is no different from traditional electricity customers (*e.g.*, commercial buildings).
- As shown in Fig. 1(b), CSPs can also perform GLB for various purposes, including but not limited to reducing the total electricity cost of its geo-distributed data centers. As long as the quality of service does not degrade, routing service requests to data centers at locations with cheaper electricity price can provide important cost reduction [26]. According to the widespread estimate in [23], the workload of a data center that can be geographically load-balanced corresponds to 20-30% of the data center electricity demand. In such scenario, CSPs represent a *new* type of electricity customers to local utilities, whose energy demand at a location is *elastic* (caused by CSPs moving their workload around).

There have been works studying the economic benefit of GLB to CSPs, under the assumption that the electricity prices seen by CSPs are not affected by GLB. However, as shown in the next section, as GLB introduces additional uncertainty in the local demand, utilities have to increase electricity prices to ensure certain profit margin in face of such demand uncertainty, cancelling the benefit of GLB. The alarming observation motivates us to consider a broker-assisted GLB solution as a clean alternative in Sec. 4.

3. ELECTRICITY PRICING MODEL

The electricity prices that CSPs pay are the result of the trading at each step of the supply chain. Any trading inefficiency along the chain reflects into the final prices. A well-known example is the extremely high electricity retail prices in California during 2001, which were due to inefficiencies coming from the spot markets [16]. Furthermore, inefficiencies may also arise between a utility and a CSP. Demand uncertainty may result in economic loss for the utility due to over-/under- estimation of the required supply.

In this section we present a model that shows how utilities have to increase retail prices in order to ensure certain profit margin in

[2]In the day-ahead market that we are interested in, electricity supply for each hour of the next day is auctioned. Without loss of generality, we focus on the auction for the electricity supply of a particular hour.

[3]Buyers that could not get their bids matched in the day-ahead market can attempt to get their supply in subsequent real-time markets. However, generating sources with short response-times, such as gas turbines, are expensive and they cannot be permanently running. As a result, the average MCP of real-time markets are likely to be more expensive and changing [2, 25].

face of the economic loss caused by GLB. Consequently, CSPs doing GLB (as in Fig. 1(b)) actually may end up paying *higher* electricity bills than not doing GLB (as in Fig. 1(a)).

3.1 Prediction Error Increases Retail Price

We begin by showing how larger errors in demand prediction will lead to higher retail prices. Utilities make profit by determining a proper retail price for selling electricity. Let d be the actual demand for a particular hour in the next day and $\tilde{d}$ be the utility's prediction of d. Let w_b be the average (MCP) price at which the utility purchased $\tilde{d}$ amount of electricity for that hour from the day-ahead market.

Without prediction error, *i.e.*, $\tilde{d} = d$, given a price[4] p_0, the utility obtains a desired expected profit for the hour as

$$(p_0 - w_b)\, d. \tag{1}$$

With prediction error, the utility suffers economic loss as compared to the error-free case.

- In case of over-prediction, there is $\tilde{d} - d > 0$ amount of electricity surplus (and it cannot be stored). In today's practice, the utility can sell them back to a GENCO at an average marginal price denoted as w_s (usually $w_b > w_s$). The economic loss to the utility is $(w_b - w_s)\left(\tilde{d} - d\right)$.

- In case of under-prediction, there is $d - \tilde{d} > 0$ amount of unmatched demand to be urgently balanced by the utility to avoid power outage. In today's practice, the utility can purchase supply in the hour-ahead or real-time markets to satisfy urgent demand, but at a price higher than in day-ahead markets. Denote the average marginal price of buying electricity in urgency as w_u ($w_u > w_b$). The economic loss to the utility is then $(w_u - w_b)\left(d - \tilde{d}\right)$.

In order to compensate the economic loss of the utility due to prediction error, and to obtain the same expected profit in Eq. 1, the utility needs to set a retail price p *higher* than p_0 (the price for the error-free case) according to:

$$\begin{aligned} p = p_0 &+ (w_b - w_s)\, \mathbb{E}\left[\left(\tilde{d} - d\right)^+ / d\right] \\ &+ (w_u - w_b)\, \mathbb{E}\left[\left(d - \tilde{d}\right)^+ / d\right] > p_0. \end{aligned} \tag{2}$$

In today's practice, prediction error is specified in terms of *mean absolute percentage error* (MAPE), defined as

$$\Delta d = \mathbb{E}\left[\left|\tilde{d} - d\right| / d\right]$$

With only MAPE available, the utility can define its price as

$$p = p_0 + (w_u - w_s)\, \Delta d. \tag{3}$$

3.2 Market Volatility Increases Retail Price

So far, we have considered that the MCP w_b is a value provided by the market. In practice, following the discussion in Sec. 2.1, this wholesale price depends on whether the bids that the utility places in the market auction are granted with supply. We model that in each location there are a day-ahead market and a real-time market, which run for each hour (of the next day). We also assume that the utility places a single bid $(b, \tilde{d})$ in the day-ahead market; here b represents the bidding price, while $\tilde{d}$ is the bidding electricity quantity, which must match the predicted demand. If the bid fails to win in the auction, i.e., it is lower than the day-ahead MCP, then the electricity is purchased at the real-time MCP.

Denote w_t and w_u as the MCPs of the day-ahead and real-time markets, respectively. Then, the marginal electricity price for a particular hour is

$$w_b = \begin{cases} w_t, & \text{if } b \geq w_t; \\ w_u, & \text{otherwise.} \end{cases} \tag{4}$$

Based on historical data, the MCP distribution for each market can be estimated. Denote f_t and f_u as the probability density functions of w_t and w_u, respectively. Then the average wholesale price $\overline{w}_b$ is given by,

$$\mathbb{E}[w_b] = \int_0^b x \cdot f_t(x)dx + \mathbb{E}[w_u] \int_b^\infty f_t(x)dx \tag{5}$$

Note the dependency on the bidding price. The right term of the equation, $\Delta w = \mathbb{E}[w_u] \int_b^\infty f_t(x)dx$, is an additional (marginal) cost due a wrong estimation of the MCP while choosing the bidding price. Following a similar development as for the demand prediction error, market volatility increases the average retail price like

$$p = p_0 + (w_u - w_s)\, \Delta d + \Delta w. \tag{6}$$

3.3 Discussion: Incentives for Coordination

Based on this electricity pricing model, demand and MCP prediction are critical for the operation of the utilities. Currently, demand prediction is negligible. Electrcity demand is rather predictable as it follows patterns that repeats daily, with seasonality during weekends and holidays.

Although its impact depends on the amount of routed electricity, GLB may introduce utterly different demand patterns. When used extensively, the difficulty for a utility to predict routed electricity demand is that the demand also depends on the prices in other locations, which may not be disclosed timely to the utility (unless the utility is a market participant in all the locations). Therefore, just by adapting local demand prediction methods to GLB may not be enough to yield accurate predictions. On the other hand, coordination does removes this demand-side uncertainty, it is possible the first incentive for the utilities to coordinate with CSPs.

The second incentive is related to Demand-Side Management (DSM) techniques. GLB and DSM are similar in the sense that they both make the demand elastic to prices. As prices increase for a particular time, users may decrease the demand. The difference resides in that, while DSM defers demand to off-peak hours *when* prices are cheaper, GLB routes demand to other locations *where* prices are cheaper.

Consequently, since demand elasticity depends on the prices in other locations, then (i) the demand may not be deferred to other hours as in conventional DSM, but consumed in other locations, possibly served by other utilities operating there and thus decreasing the local demand. (ii) Off-peak reduction may be more difficult to achieve in locations where prices are cheaper. Even if prices are increased to that end, they may still be cheaper than in other locations.

Summarizing, for utilities it is harder to deploy DSM unilaterally, having another incentive to coordinate with CSPs. In the next section, we introduce a cooperative model in which CSPs doing

[4]The process of how a utility determines its retail price can be highly involved (consideration factors include competition from other local utilities). A vital requirement that the price has to be high enough to guarantee the (expected) profit is larger than a minimum for the utility to stay in business.

GLB can exploit their positioning in multiple locations. In this scenario, if the CSP and utility cooperates, then demand uncertainty (Δd) is suppressed.

In addition, the fact that a part of the CSPs demand does not have to be attended locally, allows to purchase and consume the electricity in multiple locations. That implies having more opportunities to obtain that electricity, what immediately decreases the level of risk faced in the auctions and thus the market uncertainty (Δw).

As for the utilities, this model also present incentives to cooperate as they could retail at lower prices, avoid a potential operatonal losses when CSPs use GLB extensively and simplify their purchase of electricity in the markets, at least the supply corresponding to the datacenters.

4. A BROKER-ASSISTED GLB SOLUTION

Motivated by the recent practice of large CSPs moving into electricity markets and the deployment of *SmartGrid* infrastructure, we propose a cooperative scenario that is efficient with respect to the pricing model we just shown. In this scenario, CSPs purchase their electricity needs either directly from markets or through brokers. By doing GLB and electricity procurement jointly, CSPs can eliminate the trading inefficiencies we discussed in the previous section. Consequently, GLB can be used without causing economic losses to utilities, and CSPs can still exploit their presence in multiple geo-locations to reduce electricity cost.

Implicitly, this approach creates a "geo-distributed supply chain" illustrated in Fig. 1(c), in which a large CSPs like Google, which can buy and sell electricity directly from/to the spot markets since 2010 [14], can first bid in different spot markets, *i.e.*, *balance its bids*, and then it can balance its load across geo-distributed data centers according to the obtained electricity supply, *i.e.*, *balance its bits*.

4.1 Joint GLB and Electricity Procurement: Problem Formulation

In our broker-assisted solution, the CSP needs to solve a joint GLB and electricity procurement problem. We first present the setting and necessary notation. Without loss of generality, we consider the problem for a particular hour of a day. Consider a CSP that receives U_i amount of service requests from location i ($1 \leq i \leq m$), and it runs data centers at n locations where data center at location j has a capacity of C_j ($1 \leq j \leq n$). We assume the CSP, based on their service history, can estimate U_i ($1 \leq i \leq m$) accurately, which is aligned with the recent successes on using time series analysis for estimating user service requests [11].

We model the GLB quality of service constraint by defining $a_{ij} = 1$ if data center at location j serves requests at location i with satisfactory quality of service, $a_{ij} = 0$ otherwise.

Let z_{ij} be the corresponding network cost of serving one request from location i in the data center at location j. Let $u_{ij} \geq 0$ be the amount of requests from location i served by the data center at location j, then the total requests served by data center at location j is $\sum_{i=1}^{m} u_{ij}a_{ij}$. Let γ be the conversion ratio that maps the total requests to the amount of electricity needed to serve the requests[5], then the electricity demand for serving $\sum_{i=1}^{m} u_{ij}a_{ij}$ amount of requests is simply $\tilde{d} = \gamma \cdot \sum_{i=1}^{m} u_{ij}a_{ij}$.

We denote the expected wholesale price at location j as $\mathbb{E}\left[w_b^j\right]$. Recall this expected value is a function of the bidding price b_j, real-time expected price w_u^j and the probability density functions f_t^j and f_u^j (see Eq. 11).

[5]For example, as reported by Google [26], each search consumes 0.28 Watts-hour electricity in its data centers.

With the above notations, we can formulate the joint GLB and electricity procurement problem for a CSP as follows:

$$\min \sum_{j=1}^{n} \mathbb{E}\left[w_b^j\right] \cdot \sum_{i=1}^{m} u_{ij}a_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}u_{ij} \tag{7}$$

$$\text{s.t.} \sum_{j=1}^{n} u_{ij}a_{ij} \geq U_i,\ 1 \leq i \leq m, \tag{8}$$

$$u_{ii} \geq \alpha \cdot U_i, 1 \leq i \leq m, \tag{9}$$

$$\sum_{i=1}^{m} u_{ij}a_{ij} \leq \min\left\{C_j, \frac{1}{\gamma}\tilde{d}_j\right\},\ 1 \leq j \leq n, \tag{10}$$

$$\mathbb{E}\left[w_b^j\right] = \int_0^{b_j} x \cdot f_t^j(x)dx + \mathbb{E}\left[w_u^j\right] \int_{b_j}^{\infty} f_t^j(x)dx, \tag{11}$$

$$\text{var. } u_{ij} \geq 0, b_j \geq 0, \tilde{d}_j \geq 0, 1 \leq i \leq m, 1 \leq j \leq n.$$

In the above problem, the objective in Eq. 7 represents the total expected cost of energy procurement and network load balancing cost. The constraints in Eq. 8 say that the demand at every location i must be served. The constraints in Eq. 9 put a minimum on the percentage of the demand that must be served locally; here $0 \leq \alpha \leq 1$ is a pre-assigned constant. The constraints in Eq. 10 mean that the total allocated requests to data center at location j can exceed neither its physical capacity (*e.g.*, the number of total servers) nor the "effective" capacity determined by the purchased electricity $\tilde{d}_j$. Eq. 11 is the closed-form formula of $\mathbb{E}\left[w_b^j\right]$ expressed in b_j, f_t^j, and f_u^j (since $\mathbb{E}\left[w_u^j\right] = \int_0^{\infty} x \cdot f_u^j(x)dx$).

We propose the following algorithm to find the optimum value of this formulation.

4.2 An optimal algorithm

The approach we follow to solve the joint GLB and electricity procurement problem in Eq. 7-11 is based on the following observation. If $\left(b_j, \tilde{d}_j\right)$ $(1 \leq j \leq n)$ are given, then $\mathbb{E}\left[w_b^j\right]$ and $\min\left\{C_j, \frac{1}{\gamma}\tilde{d}_j\right\}$ are fixed and the above problem in Eqs. 7-11 reduces to the standard GLB formulation in [26].

Thus the novelty of our problem resides in that the electricity must be obtained from the auctions, so that the electricity price and quantity are now random variables. Consequently, the GLB optimization becomes a stochastic optimization problem. So as to solve this problem, we use a two-stage technique used in stochastic programming. In our case, this means that first we try to obtain the optimal bids, denoted as $\left(b_j^*, d_j^*\right)$ $(1 \leq j \leq n)$, and then in a second stage, once the outcome of the auctions is known, we solve the remaining GLB problem given $\left(b_j^*, d_j^*\right)$ $(1 \leq j \leq n)$.

The first difficulty we find is that in our case the first stage is a non-convex problem, in general. The term $E[w_b^j]$ in the objective function is not convex in b_j due to the arbitrary form of $f_t^j(x)$. Although non-convex problems are difficult to solve and there are no standard techniques for solving them, we can solve the problem in Eq. 7-11 in a divide-and-conquer manner. Note that, the optimal bidding price b_j^* does not depend on the workload assignment u_{ij}.

Based on this observation, we can solve the non-convex problem sequentially. Hence, we have two stages and for the first stage we need to solve two sub-problems, **SP1-j** and **SP2**. Problem **SP1-j** provides the optimum bidding price b_j^* by minimizing the price expectation in each regional market. Afterwards, we use the outcome of **SP1-j** in **SP2** and we compute the bid quantities d_j^*.

Therefore, for any $1 \leq j \leq n$ we solve **SP1-j**

$$\textbf{SP1-j}: \quad \min_{b_j} \quad E[w_b^j].$$

Even though **SP1-j** is still a non-convex problem, it reduces to compute the optimal bid in an auction with substitute items (all valued at the MCP). The optimal bidding strategy in this type of auctions is to bid at the *true price*. In our case in which we assume that the utility aims at profit zero, the true value is directly $b_j^* = E[w_u^j]$. That is because the CSP will always buy the electricity at the real-time (expected) price, as long as the bid fails in the day-ahead auction. Hence, the true price is the price expectation in the local real-time market.

Second, let the minimum expected price in **SP1-j** be $(w_b^j)^*$, then we solve **SP2** to determine the tentative workload assignment u_{ij}^* in the first stage.

$$\textbf{SP2}: \quad \min \sum_{j=1}^{n} (w_b^j)^* \cdot \left[\sum_{i=1}^{m} u_{ij} a_{ij}\right] + \sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij} u_{ij}$$

$$\text{s.t.} \sum_{j=1}^{n} u_{ij} a_{ij} \geq U_i,\ 1 \leq i \leq m,$$
$$u_{ii} \geq \alpha \cdot U_i, 1 \leq i \leq m,$$
$$\sum_{i=1}^{m} u_{ij} a_{ij} \leq C_j,\ 1 \leq j \leq n,$$
$$\text{var. } u_{ij} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n.$$

SP2 is a linear programming problem and we can solve it to get its optimal solution u_{ij}^* by standard techniques. After that, we can calculate the bidding quantity d_j^* for the j^{th} data center by

$$d_j^* = \gamma \sum_{i=1}^{m} u_{ij}^* a_{ij}, 1 \leq j \leq n. \tag{12}$$

After the auctions are executed, depending on the outcome and the winning and non-winning bids, we get the final electricity price w_b^j and amount d_j. These values are used to solve the second and final stage, from which we obtain the final workload assignment u_{ij}. This workload is computed by replacing the expected values, $\mathbb{E}\left[w_b^j\right]$ and $\tilde{d}_j$, in Eqs. 7-11 by w_j and d_j.

Algorithm 1 Algorithm for *Broker-assisted GLB*

1: **for** $1 \leq j \leq n$ **do**
2: $b_j^* \leftarrow E\left[w_u^j\right]$;
3: $E[w_b^j]^* \leftarrow \int_0^{b_j^*} x \cdot f_t^j(x)dx + \mathbb{E}\left[w_u^j\right] \int_{b_j^*}^{\infty} f_t^j(x)dx$;
4: **end for**
5: Get the (optimal) tentative workload assignment $[u_{ij}^*]_{n\times m}$ by solving **SP2**;
6: **for** $1 \leq j \leq n$ **do**
7: $d_j^* \leftarrow \gamma \sum_i u_{ij}^* a_{ij}$;
8: bid with (b_j^*, d_j^*);
9: **end for**
10: Do conventional GLB with auction outcomes w_b^j and d_j to obtain the final optimal workload assignment u_{ij}

We summarize the above understandings into *Algorithm* 1 and show its optimality by *Theorem 1*. For $1 \leq i \leq m, 1 \leq j \leq n$,

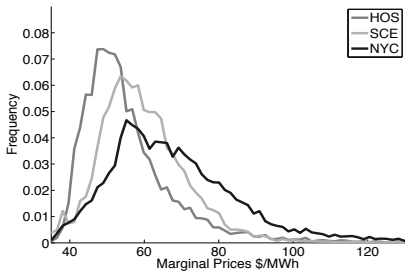


Figure 2: MCP distribution for San Diego (SCE), Houston (HOS) and New York City (NYC) in 2009 - 2012

Theorem 1. *The bids (b_j^*, d_j^*) and workload assignment u_{ij} obtained by Algorithm 1 are an optimal solution to the problem in Eq. 7-11.*

PROOF. see *Appendix 1* □

5. PERFORMANCE EVALUATION

We evaluate the performance of the broker-assisted GLB solution that we propose. First, we describe and characterize the dataset we use in the experiments. Part of this analysis serves as a validation of the incentive analysis and assumptions made about the workload. Second, we analyze the cost effectiveness and optimality of the solution.

Our evaluation is carried out in a scenario in which a (virtual) CSP operates three data centers, in San Diego, Houston, and New York City. We choose this scenario as it reflects the tendency of large CSPs, such as Google and Facebook, to deploy customized data centers in the East, Mid, and West part of the US. For this scenario we take into account the following considerations:

Ratio of workload eligible for GLB: As Eq. 9 shows, some portion of workload must be served locally while the rest is *eligible* for GLB. We denote the percentage of the workload that is eligible for GLB as β, and $\beta = 1 - \alpha$. The proportion between them is usually due to many factors that vary from one provider to another, e.g., response time requirement, resource and information availability, SLAs, etc. Therefore, we investigate the performances of our Broker-assisted GLB with different values of β.

Topology constraints: The same as in [26], we assume that eligible workload is re-routed maximum between two consecutive locations before seemingly degrading the quality of service. Therefore, the CSP can balance San Diego's eligible load between San Diego and Houston, New York's load between New York and Houston, and Houston's among the three locations.

Internet-related costs: Furthermore, in our simulation we do not consider the internet cost. It does not mean that we treat that kind of cost as negligible, yet we adopt this model as (i) large internet companies like Facebook or Google negotiate with carriers on a nationwide basis, therefore bandwidth prices are usually not geographically differentiated [26]; (ii) Internet transit costs are referred the maximum traffic during a time period or the so called 95-percentile of the bandwidth allocation [7]. Thus we assume that network related cost are similar regardless the use of GLB.

Data centers' capacity: The maximum workload that CSPs can re-route is estimated nowadays between $20 - 30\%$. Taking into account that not all load is eligible for GLB, we assume that data centers' capacity is large enough to cope with additional incoming load due to GLB. We justify this assumption using the fact that typically data center owners overdimension their capacity, typically, at least by 20% [1].

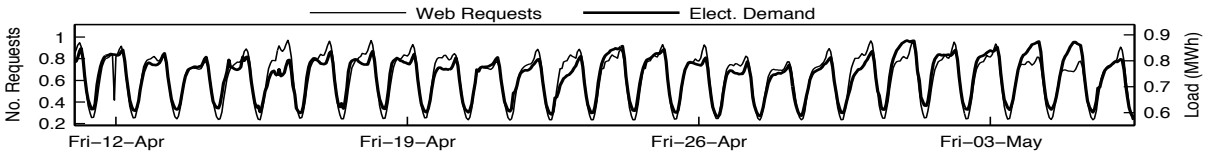


Figure 3: Evolution of the (aggregated) electricity demand and web workload between April 12th and May 6th 2013.

5.1 Dataset Characterization

Electricity demand and prices: To obtain the total electricity demand for each of the three local utilities, we crawl the hourly electricity demand from the spot markets in San Diego [10], CA, Houston [12], TX, and New York [24], NY for 2009-2012, and choose nodal demand until the data center demand represents to 30% of the utility's demand (following the back-of-the-envelope computation presented in the introduction). We also collect the hourly MCPs of the three spot markets for the same period. The empirical distributions of the MCP for the three markets are shown in Fig. 2.

Finally, to maximize their prediction accuracy, utilities take into account the weather conditions and daily activity patterns. We crawl the hourly weather conditions [35] in the three areas and the official holidays calendar for 2009 - 2012. We omit the weekends in all our experiments, due to the seasonality of the workload and electricity demand during these days.

CSP Workload: We use traces from the Akamai CDN as the user request workload of the (virtual) CSP in its three data centers. We crawl Akamai's Internet Observatory website [8] to obtain the number of HTTP requests per minute against the Akamai CDN in North America. Akamai CDN relies on co-location data centers that individually do not represent large electricity consumption. Nevertheless, using the conversion rate of $1kJ$ per query (0.28 $Watts \cdot h$) claimed by Google for its data centers [26], the crawled workload aggregately creates a power consumption of 125 MW, which may serve well to approximate the consumption of three Facebook's data centers at full utilization (according to [5, 3]).

Since Akamai does not dissect the information of its workload per location, we have run a preliminary experiment to make an educated approximation of the workload splitting for the three locations. We aggregate the electricity demand curves from the three locations into a time series, respecting the time difference between the aggregated time series of each location. We compare this (normalized) electricity demand aggregate with the time series of the (normalized) number of web requests against the Akamai CDN. The two series are displayed in Fig. 3.

The correlation coefficient of these aggregated curves is 0.92. Most differences appear during the morning and more noticeably in some weekends, what we associate with the industrial and commercial activity. If we take into account that the three areas we are using have similar development levels, then it is reasonable to assume that a random sample among the population of these three areas will provide similar results about the usage of electricity and web services (and the ratio between these two). Therefore, splitting the number web requests in each location according to the ratios of electricity demand among the locations should provide us with a good approximation.

5.2 GLB increases Utilities Prediction Error

As GLB dynamically allocates energy-intensive workload to data centers at different geo-locations, it increases electricity demand variation for the local utilities. According to Eq. 6, if this variation introduces prediction errors, it may result in higher retail prices. To assess such phenomenon, we carry out a case study based on our real-world dataset and analyze to what extent this extra demand variation will lead to larger errors in utilities' demand prediction.

Table 1: MAPE and Prices vs. Balanced Load

GLB	San Diego		Houston		New York	
(%Load)	MAPE (%) & Avg. Price ($/MWh)					
0	3.0	47.9	2.7	43.9	3.0	70.2
15	6.8	49.3	3.5	45.5	6.4	70.8
30	8.2	49.8	7.3	47.2	7.6	71.0
45	10.7	50.8	10.5	48.7	8.6	71.2
60	14.3	52.2	14.8	50.8	10.7	71.6
MAPE/GLB	0.714		0.921		0.345	

We evaluate the prediction error of the utility. We change the demand corresponding to the allowed GLB workload between 0-30% of the total utility demand. We also extend the range up to 60% to evaluate a futuristic scenario reflecting the data center electricity demand growth. For each hour, the CSP solves a standard GLB cost-minimization problem as the one in [26] to allocate its allowed GLB workload optimally.

The evaluation is carried out assuming that utilities use commonly adopted *neural networks* (NN)-based demand forecast algorithms [32] to predict their electricity demand[6]. Utilities use NNs as a black-box, which require training with sample data. Once they are trained, for each hour, the NN takes as inputs the weather forecast, historical demand records, and whether it is a public holiday/weekend or not. Based on these input values, the NN predicts the demand for that particular hour, with a certain estimation error.

We train the NN with data from 2009-2011 and use the trained algorithm to perform hourly demand prediction during 2012. To this end, we use different training datasets, one for the case without GLB and one for each GLB eligible ratio that we study (for that we perform GLB on the training load as well). We compare the prediction and the actual demand, record the MAPE, and compute the retail prices with and without prediction errors according to Eqs. 1 and 3 with $p_0 = w_b$ (modeling an altruistic utility targeting zero expected profit in the error-free case).

Electricity Demand Prediction Error: The results of how GLB affects retail prices are summarized in Table 1. Each data center location has two associated columns. The first column shows the MAPE in the presence of varying GLB load (in percentage, increased at 15% resolution). The second column is the corresponding average retail prices according to Eq. 3. The last row shows the ratio MAPE per % of routable load to other locations.

Several interesting observations can be made. First, without GLB (corresponding to the third row of 0% GLB load), the NN algorithm can predict the actual demand pretty accurately – with a MAPE at most 3%. A closer look into the prediction accuracy of the NN al-

[6]For a real case, see `http://www.mathworks.com/tagteam/63938_91460v01_GasNaturalFenosa_English_final.pdf`.

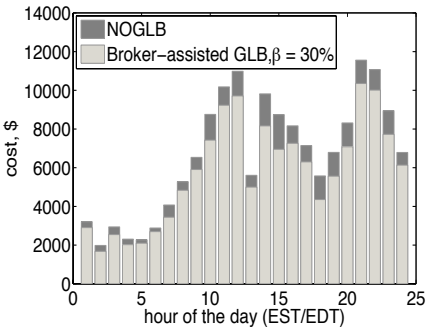
gorithm for the San Diego site shows the hourly MAPE has a mean of 3% and a standard variation of 6%. These results show that without GLB, NNs can predict accurately the real-world electricity demand, justifying its widespread adoption in practice.

Second, as the GLB load percentage increases, MAPE of the NN algorithm becomes worse. For example, in Table 1, when the GLB load increases to 30%, the MAPE for San Diego increases to 8.15%, 2.7 times of that of no GLB. The standard deviation of MAPE is 11.3%, almost twice of that of no GLB. These results are in sharp contrast to the case of no GLB, and confirm our intuition that GLB introduces demand uncertainty and extra errors in the demand prediction.

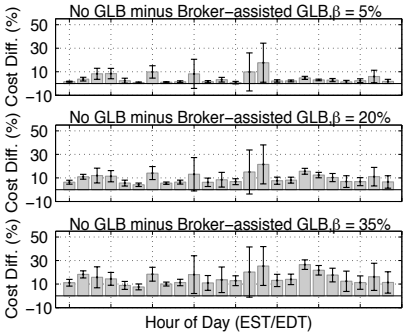
Increment on the Electricity Bill for CSPs: Using the pricing model from Section 3, we can compute the increment in the retail prices corresponding to the economic loss of the utility. This is displayed also in Table 1; **the retail price for San Diego on average increases by 0.7% for every increment of 1% in the GLB load.**

We add to the pricing information the workload allocation to compute the cost. We do it for the cases where the CSP is able to move $\beta = 0\%, 15\%, 30\%$, and 60% of the total local utility demand. We study and compare the total electricity cost (sum of the three locations for the year 2012) between the baseline case, $\beta = 0\%$, and the rest (in percentage).

Results show that in the $\beta = 15\%$ case **the CSP actually ends up paying a total bill 1% higher than not doing GLB at all**. In the $\beta = 30\%$ case where the CSP can move up to 30% of its overall workload, the ability to aggressively move workload to low-price locations improves the results, despite the increase in the electricity prices due to higher degrees of uncertainty. However, there is still minor savings in the overall electricity bill, about 3%, while the CSP is already moving the full allowed GLB workload of its data centers. Finally, higher benefits could be achieved for large *allowed* GLB load. For the $\beta = 60\%$ case, the GLB effect provides 9% cost reduction. However, this case requires the CSP to move a workload that is beyond the *feasible* percentage in data centers nowadays (20-30% [26]).



(a) Cost comparison between the $\beta = 0\%$ and broker-assisted with $\beta = 30\%$ case.



(b) Hourly cost difference between $\beta = 0\%$ and the three *Broker* cases, $\beta = 5\%$, 20% and 35% GLB.

Figure 4: CSP cost for each hour of the day as 4(a) average cost and 4(b) differential w.r.t. the $\beta = 0\%$ case.

5.3 Cost Reduction with Joint Procurement

We evaluate the performance of the Algorithm 1 in the scenario with three data-centers. We have implemented the algorithm in a simulator of market auctions, which we feed with our dataset.

We first assume that 30% of the workload can be routed freely for the broker-assisted GLB model, i.e. $\beta = 30\%$, and compute the average cost per hour for the CSP with or without GLB. The result is displayed in Fig. 4(a). Each bar represents, for each hour of the day, the average cost. The results show the day and night pattern of the web requests. From the figure we also identify two valleys at noon and around 5-7 pm, which we associate with lunch time and commuting after work and the peak hour around 8-9 pm.

The gray portion of the bars are the average cost for our broker-assisted solution. The darker portion of the bars represent the extra cost in average that the CSP pays in the $\beta = 0\%$ case, which is noticeable in most of the hours. The total electricity can be reduced by 12.8% in average on a daily basis.

Results on a yearly basis are shown in Fig. 5 (real workload curve). The tendency shows that larger values of β lead to higher cost reduction remains. The cost reduction ratio is 4.13% when $\beta = 5\%$ and increases steadily to 12.7% for $\beta = 30\%$.

5.4 Performance degradation by workload prediction error

One of the main assumptions behind the suitability of broker-assisted GLB solution is that, as a consequence of their geographic diversification and ability to accommodate demand, CSPs do a more (economically) efficient use of the electricity supply-chain. Whether this assumption holds depends strongly on the ability of CSPs to predict workloads more efficiently than how utilities predict electricity demand.

In the simulations shown until now, we assumed that CSPs know the future workload exactly. We now re-compute the cost reduction ratio by executing Alg. 1 over predicted workload instead of real workload. For the predicted workload case, we sample the workload of the same hour and same day of the week in the previous 5 weeks and use the sample average as estimator. Results displayed in Fig. 5 show the impact of the workload prediction error. The cost reduction ratio is lower for all values of β when the predicted workload is used instead.

We also test the broker-assisted GLB's performance with different prediction errors (MAPE), while setting $\beta = 20\%$. Consistently, the curve in Fig. 6 shows that the bigger the prediction error, the more the performance degrades.

Finally, after showing that the workload MAPE does have an impact on the cost reduction ratio, we compare the CSPs' workload MAPE with the utilities' electricity demand MAPE. The prediction error distribution for each hour of the day during 2012 are shown for the workload MAPE in Fig. 7(a) and for the electricity demand MAPE in the San Diego market, cases $\beta = 0\%$ and $\beta = 30\%$, in Figs. 7(b)-7(c).

Although the prediction method used for the CSP workload, described above, is less elaborated than the NN we used for the utili-

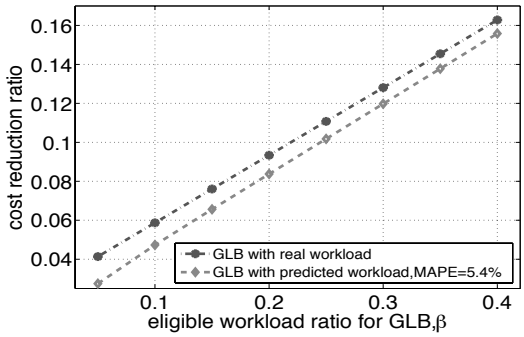


Figure 5: cost reduction ratio for GLB with real workload and predicted workload, as β increases from 5% to 40%

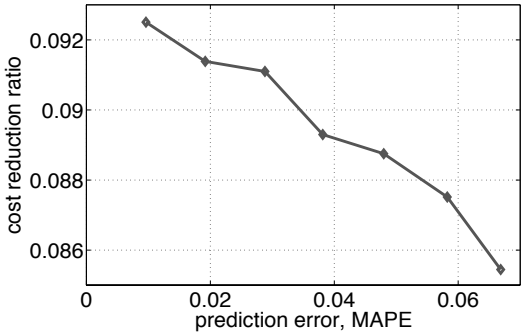


Figure 6: cost reduction ratio with different prediction errors, as $\beta = 20\%$

ties' demand prediction, it achieves only 5% MAPE, which is close to the 3% of the $\beta = 0\%$ and substantially more efficient than the MAPE for the $\beta = 30\%$ case. This results point out that if GLB is used, CSPs can do a more effective prediction of their demand than the utilities.

6. RELATED WORK

The seminal work suggesting the use of GLB to reduce the electricity bill of geo-distributed data-center owners is probably by Qureshi *et al.* [26]. Subsequent publications analyzed the technical feasibility and assess the possibilities of GLB [20, 36, 33, 22, 13]. All these works consider that GLB is innocuous to the electricity prices, what we show it is a strong assumption. We suggest that our broker-assisted model opens the possibility to exploit the advantages of these works without any undesired effects for the utility companies. More recently, other works start considering the potential of using spot markets information in data centers [17]. They show promising results, using markets information to defer energy consuming tasks in data centers while elevated prices are accused. Compared to our broker-assisted solution, they do not explore the benefits of a jointly scheduling of energy purchase and consumption.

Regarding this optimal procurement, cloud-providers are completely new players in the electricity markets. In fact, pricing models specific to datacenter demand response has been recently proposed [21]. These pricing models analyze the demand response of only one datacenter. We consider several datacenters instead, showing that in contrast to the utilities, the CSP is able to bid more efficiently in markets in different locations. Finally, the optimization of those bids also provides a novel study case for the existing literature on strategic bidding [29, 15, 19].

7. CONCLUSIONS

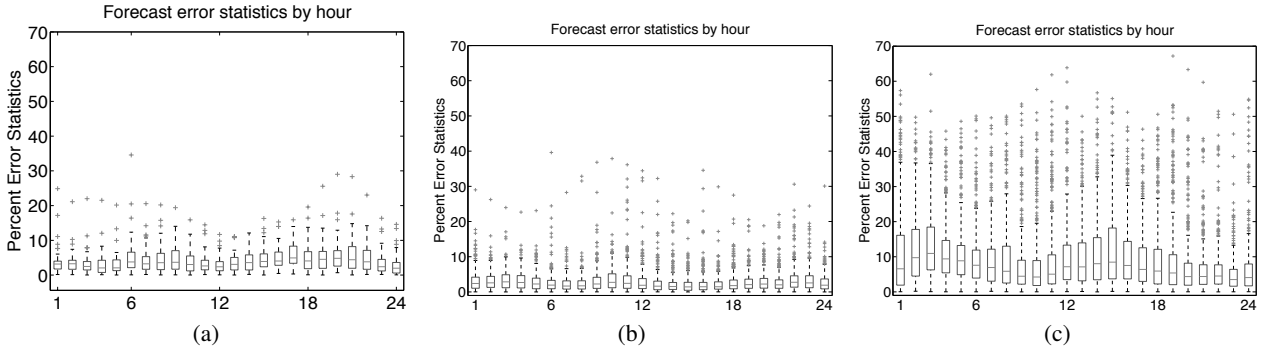
We carry out a comprehensive study of the potential of GLB on reducing the electricity bills for CSPs that operate multiple geo-distributed data centers. By analysis and case study using real-world traces, we show that as GLB introduces extra uncertainty in local electricity demand, it causes trading inefficiency between local utilities and CSPs and subsequently economic loss to the utilities. As such, to ensure certain profit margin in face of such GLB-induced economic loss, utilities will have to increase electricity prices. This challenges the common assumption in existing studies that GLB has no impact on electricity prices. Our study reveals a perhaps surprising observation – CSPs doing GLB can see poor cost reduction or even pay more in electricity than not doing GLB. We then propose to allow CSPs to purchase electricity from markets through brokers. By doing GLB and electricity procurement jointly, CSPs eliminate the trading inefficiency between them and utilities and the economic loss to utilities. Meanwhile, CSPs can still exploit their presence in multiple geo-locations to reduce electricity bills – up to 12% less than not doing GLB, for our case study based on real-world traces.

8. ACKNOWLEDGEMENTS

This work was partially supported by National Basic Research Program of China (Project No. 2012CB315904 and 2013CB336700) and several grants from the University Grants Committee of the Hong Kong S.A.R., China (Area of Excellence Grant Project No. AoE/E-02/08, Theme-based Research Scheme Project No. T23-407/13-N, and General Research Fund Project No. 411011).

9. REFERENCES

[1] Data center knowledge archive. available at `http://www.datacenterknowledge.com`.
[2] Duke energy annual report 2011. `http://www.duke-energy.com/pdfs/DukeEnergy_2011_AR-10k.pdf`.
[3] Facebook's new 'cloud'. Technical report, ECONorthWest, 2011.
[4] 2012 state of the markets report. Technical report, Federal Energy Regulatory Commission, 2012.
[5] How clean is your cloud? Technical report, Greenpeace Climate, 2012.
[6] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya. A taxonomy and survey of energy-efficient data centers and cloud computing systems. *Advances in Computers, vol. 82, pp. 47-111*, 2011.
[7] M. Adler, R. K. Sitaraman, and H. Venkataramani. Algorithms for optimizing the bandwidth cost of content delivery. *Computer Networks*, 55(18):4007–4020, December 2011.
[8] Akamai Internet Observatory website.
[9] L. Barroso and U. Holzle. The case for energy-proportional computing. *IEEE Computer*, 40(12):33–37, 2007.
[10] Caiso archive. available at `http://www.caiso.com`.
[11] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao. Energy-aware server provisioning and load



(a) (b) (c)

Figure 7: (a) MAPE for the CSP workload prediction error; (b) MAPE for the utility electricity demand prediction without GLB; (c) MAPE for the utility electricity demand prediction with GLB at 30%.

dispatching for connection-intensive internet services. In *Proc. USENIX NSDI*, 2008.

[12] Ercot archive. available at `http://www.ercot.com`.

[13] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav. It's not easy being green. In *Proceedings of the ACM SIGCOMM 2012*, pages 211–222, New York, NY, USA, 2012. ACM.

[14] Google energy wiki. `http://en.wikipedia.org/wiki/Google\_Energy`.

[15] R. Herranz, A. Munoz San Roque, J. Villar, and F. Campos. Optimal demand-side bidding strategies in electricity spot markets. *Power Systems, IEEE Transactions on*, 27(3):1204 –1213, aug. 2012.

[16] P. Joskow. California's electricity crisis. *Oxford Review of Economic Policy*, 17(3):365–388, 2001.

[17] C. Kelly, A. Ruzzelli, and E. Mangina. Using Electricity Market Analytics to Reduce Cost and Environmental Impact. In *Proceedings of the 2013 IEEE Green Technologies Conference*, pages 414–421, 2013.

[18] J. G. Koomey. Growth in data center electricity use 2005 to 2010. *Oakland, CA: Analytics Press*, 2010.

[19] M. Liu and F. Wu. Risk management in a competitive electricity market. *International Journal of Electrical Power & Energy Systems*, 29(9):690–697, 2007.

[20] Z. Liu, M. Lin, A. Wierman, S. Low, and L. Andrew. Greening geographical load balancing. In *Proc. ACM SIGMETRICS*, pages 233–244, 2011.

[21] Z. Liu, I. Liu, S. Low, and A. Wierman. Pricing data center demand response. In *Proc. ACM SIGMETRICS '14*, Jun. 2014.

[22] J. Luo, L. Rao, and X. Liu. Data center energy cost minimization: a spatio-temporal scheduling approach. In *Proc. IEEE INFOCOM*, 2013.

[23] D. Meisner, B. Gold, and T. Wenisch. Powernap: eliminating server idle power. *ACM SIGPLAN Notices*, 2009.

[24] Nyiso archive. available at `http://www.nyiso.com`.

[25] 2011 Oregon Utility Statistics.

[26] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs. Cutting the electric bill for internet-scale systems. In *Proc. ACM SIGCOMM*, pages 123–134, 2009.

[27] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu. No power struggles: Coordinated multi-level power management for the data center. In *ACM SIGARCH*, volume 36, pages 48–59, 2008.

[28] N. Rasmussen. Electrical efficiency modeling of data centers. *Technical Report White Paper*, 113, 2007.

[29] S. Sethi, H. Yan, J. Yan, and H. Zhang. An analysis of staged purchases in deregulated time-sequential electricity markets. *Journal of Industrial and Management Optimization*, 1(4):443–463, 2005.

[30] R. Sharma, C. Bash, C. Patel, R. Friedrich, and J. Chase. Balance of power: Dynamic thermal management for internet data centers. *IEEE Internet Computing*, 2005.

[31] Spain energy consumption. `http://www.nationmaster.com/country/sp-spain/ene-energy`.

[32] G. K. Tso and K. K. Yau. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9):1761–1768, 2007.

[33] R. Urgaonkar, B. Urgaonkar, M. Neely, and A. Sivasubramaniam. Optimal power cost management using stored energy in data centers. In *Proc. ACM SIGMETRICS*, pages 221–232, 2011.

[34] U.S. Environmental Protection Agency. Epa report on server and data center energy efficiency. *ENERGY STAR Program*, 2007.

[35] Weatherunderground. `http://www.wunderground.com`.

[36] P. Wendell, J. Jiang, M. Freedman, and J. Rexford. Donar: decentralized server selection for cloud services. In *Proc. ACM SIGCOMM*, volume 40, pages 231–242, 2010.

APPENDIX

Proof of *Theorem 1*

PROOF. There are three different variables b_j^*, d_j^* and u_{ij} for the Problem in Eq. 7-11, with $1 \leq i \leq m, 1 \leq j \leq n$. But since the definitive allocation u_{ij} is calculated in the second stage as a correction of the first stage, we need to show optimality of the cost expectation determined by b_j^* and d_j^*. We use the following notation

Notation	Description
$\mathbf{B} = [b_j]$	bidding vector
$\mathbf{U} = [u_{ij}]$	workload assignment matrix
$\mathbb{E}[w_b^j]$	electricity price expectation with b_j
$\mathbf{P} = [z_{ij} + \mathbb{E}[w_b^j]a_{ij}]$	price matrix for workload from the i^{th} location to the j^{th} location
$\mathcal{C}(\mathbf{B}, \mathbf{U})$	cost expectation with $\mathbf{B}$ and $\mathbf{U}$

By Eq. 7, we can obtain

$$\mathcal{C}(\mathbf{B},\mathbf{U}) = \sum_{j=1}^{n}\mathbb{E}[w_b^j]\cdot\left[\sum_{i=1}^{m}u_{ij}a_{ij}\right] + \sum_{i=1}^{m}\sum_{j=1}^{n}z_{ij}u_{ij} \quad (13)$$

$$= \mathbf{U}\bullet[z_{ij}+\mathbb{E}[w_b^j]a_{ij}] \quad (14)$$

$$= \mathbf{U}\bullet\mathbf{P} \quad (15)$$

Suppose the solution by *Algorithm 1* is $\mathbf{B}^*$, $\mathbf{U}^*$, and the corresponding price matrix is $\mathbf{P}^*$. Let $\tilde{\mathbf{B}}$,$\tilde{\mathbf{U}}$ be another feasible solution and the corresponding price matrix is $\tilde{\mathbf{P}}$.

In fact, the optimal solution to **SP1** is $b_j^* = \mathbb{E}[w_u^j]$, which can be verified by

$$\begin{cases}\frac{d\mathbb{E}[w_b^j]}{db_j} \le 0, \text{when } b_j < \mathbb{E}[w_u^j]\\ \frac{d\mathbb{E}[w_b^j]}{db_j} \ge 0, \text{when } b_j > \mathbb{E}[w_u^j]\end{cases}$$

Since we obtain $\mathbf{B}^*$ by minimizing the electricity price expectation $\mathbb{E}[w_b^j]$ (**SP1**), and z_{ij} is constant, we can get

$$\mathbf{P}_{ij}^* \le \tilde{\mathbf{P}}_{ij}, \forall i,j$$

Then

$$\tilde{\mathbf{U}}\bullet\mathbf{P}^* \le \tilde{\mathbf{U}}\bullet\tilde{\mathbf{P}} \quad (16)$$

Furthermore, we obtain $\mathbf{U}^*$ by minimizing $\mathbf{U}\bullet\mathbf{P}^*$(**SP2**), i.e.

$$\mathbf{U}^*\bullet\mathbf{P}^* \le \tilde{\mathbf{U}}\bullet\mathbf{P}^* \quad (17)$$

With Eq. 16 and Eq. 17, we can get $\mathcal{C}(\mathbf{B}^*,\mathbf{U}^*) \le \mathcal{C}(\tilde{\mathbf{B}},\tilde{\mathbf{U}})$.

□