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

Jose M. Camacho[†], Minghua Chen[‡], and Dah Ming Chiu[‡]
[†] Universidad Carlos III de Madrid, Spain
[‡] Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong

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 data centers 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 6-12% of the electricity cost.**

I. 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 [1], almost enough to power the entire Spain [2]. The corresponding worldwide data center annual electricity bill is around 16 billion US dollars [1]. Today, energy cost represents a large fraction of the data center operating expense [3], and the cost is increasing at an alarming rate of 12% annually [4]. 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 [5], [6], [7] and a recent survey in [8]. 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.*, [9], [10], [11], [12] 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.

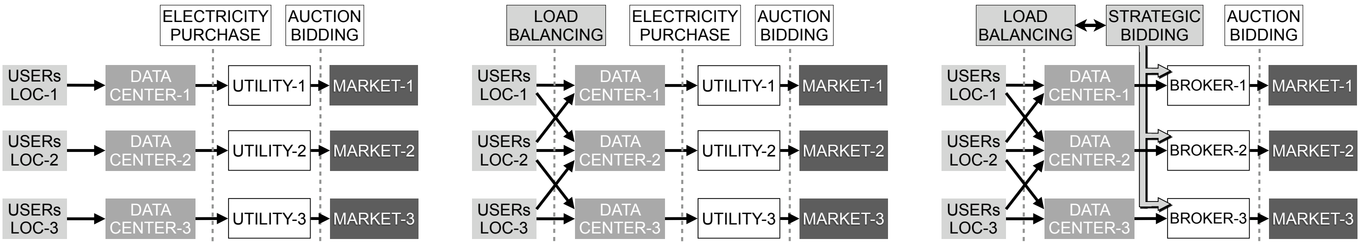
Nevertheless, all existing works focus on addressing technical feasibility and revealing the abundant benefits of GLB, assuming the electricity prices are not affected by GLB, even though GLB increases local electricity demand variation.

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 in Sec. III, prediction errors will lead to over-/under- supply and consequently 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 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 [13]. Power Pacific, a large utility serving Oregon including Prineville, sells 35 GWh daily [14]. 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 [1] while conventional demand remains steady. If data centers can shift 30% electricity demand away by doing GLB according to the estimate in [11], then GLB could lead to 10% demand variation for Power Pacific in 4 years.

Motivated by the above observations, we develop relevant



(a) Conventional electricity supply chain.

(b) Electricity supply chain with GLB.

(c) Geo-Distributed electricity supply chain.

Fig. 1. Three electricity ecosystems studied in this paper.

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. II). They are very different from conventional electricity customers whose demands are localized and inelastic.

▷ By analysis and case study using real-world traces, we investigate the interaction of GLB with the supply chain and its economic consequence (Sec. III). 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.

▷ 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. IV). 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*. Our case study using real-world traces shows that the solution can save CSPs 6-12% of the electricity cost as compared to not doing GLB.

After discussing the related work in Sec.V, we conclude the paper in Sec.VI.

II. 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 [11].

A. 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 [16].

In every spot market, the electricity supply is auctioned[2]. The *sellers*, *i.e.*, GENCOs, submit (hourly) generation offers,

[1] As a real-world example, in February 2010 the Federal Energy Regulatory Commission authorized Google to buy and sell energy at market rates [15].

[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.

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 (resp. bids), the GENCOs (resp. Utilities) specify the amount of electricity they want to sell (resp. 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 [17], [18] 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 (resp. buying bids) fail as their marginal prices are above (resp. 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 can 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*. Fig. 3(a) shows the empirical distribution of MCP (ranging from 35 $/MWh to 130 $/MWh) for three day-ahead spot markets in the US.

B. 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 [11].

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 supply 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 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 exact 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 or unused electricity, which 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. III.

C. 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 [11]. According to the widespread estimate in [20], 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. IV.

III. GLB Increases the Electricity Bills of CSPs

The electricity prices that CSPs pay are the result of the trading at each step of the supply chain. It turns out that any trading inefficiency in one step of the chain, such as the cost caused by over-supply or under-supply, propagates and reflects into the final prices. A well-known example is the extremely high electricity retail prices in California during 2001 due to inefficiencies in the spot markets [18].

[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 [19], [14].

Inefficiency may arise also between utility and CSP. As discussed in Secs. II-B and II-C, as GLB introduces demand uncertainty, it causes additional error in utility's demand prediction, resulting in economic loss for the utility due to over-/under- supply.

In this section, by analysis and case study based on real-world traces, we show that utilities will have to increase retail prices to ensure certain profit margin in face of the economic loss caused by GLB. Consequently, CSPs doing GLB (as in Fig. 1(b)) actually end up paying *higher* electricity bills than not doing GLB (as in Fig. 1(a)).

A. 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 $(w_u - w_b)\left(d - \tilde{d}\right)$.

Overall, the total expected economic loss of the utility due to prediction error is given by

$$(w_b - w_s)\,\mathbb{E}\left[\left(\tilde{d} - d\right)^+\right] + (w_u - w_b)\,\mathbb{E}\left[\left(d - \tilde{d}\right)^+\right] > 0.$$

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

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

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

$$\mathbb{E}\left[\left|\tilde{d} - d\right| / d\right] = \mathbb{E}\left[\left(\tilde{d} - d\right)^+ / d\right] + \mathbb{E}\left[\left(d - \tilde{d}\right)^+ / d\right].$$

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

$$p = p_0 + (w_u - w_s)\,\mathbb{E}\left[\left|\tilde{d} - d\right| / d\right] > p_0, \tag{3}$$

to ensure the expected profit is at least the desired one in Eq. 1.

B. GLB Aggravates 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. We analyze to what extent this extra demand variation will lead to larger errors in utilities' demand prediction, what according to our analysis in the previous subsection will result in higher retail prices. To demonstrate such phenomenon, we carry out a case study based on a real-world dataset from electricity markets and Internet web services.

Scenario: The scenario that we evaluate tries to reflect 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. In particular, in this scenario we mimic a (virtual) CSP that operates three data centers in San Diego, Houston, and New York City. The corresponding supply chains are as illustrated in Fig. 1(b): the CSP performs GLB among these three data centers, the three utilities (one at each location) buy electricity from the local spot markets based on their demand prediction and sell electricity to their local customers including the three data centers.

We assume that with quality of service considerations, the CSP can balance its San Diego load between San Diego and Houston, its New York load between New York and Houston, and its Houston load among all three locations.

From the viewpoint of three local utilities serving the data centers, the action of the GLB is perceived as demand variation. Based on the back-of-the-envelope calculation that we presented in the introduction, we assume that the data center consumption represents up to 30% of the local utility demand. If the allowed GLB workload represents 15-30% of the total CSP consumption as estimated in [11], then the GLB-induced demand variation corresponds to 5-10% of the utility's total electricity demand.

Each of the three spot markets serves a large number of local utilities. Hence, the GLB-induced demand variation (only 10% to a single local utility) has negligible impact on their MCPs, considering the vast volume of electricity trading in the markets.

Dataset:

[4] The process of how a utility determines its retail price can be highly involved (consideration factors include competition from other 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.

In order to obtain the total electricity demand of each of the three local utilities, we crawl the hourly electricity demand from the spot markets in San Diego [24], CA, Houston [25], TX, and New York [26], NY for 2009-2012, and scale them down so that the data center demand represents to 30% of the utility's total demand. 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. 3(a).

Finally, to maximize their prediction accuracy, utilities take into account the weather conditions and daily activity patterns. We crawl the hourly weather conditions [27] 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.

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 [21] 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 [11], 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 [13], [22]).

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 aggregate with the time series of the Akamai's workload. The comparison is displayed in Fig. 2, where the horizontal axis represents the time (in hours), the left y-axis the (normalized) number of web requests against the Akamai CDN and the right y-axis the (normalized) electricity demand. . The correlation coefficient is 0.92 between the aggregated electricity demand and the web request workload curves. Day and night patterns can be observed as well as weekdays and weekends. The electricity demand curve differs from the workload more noticeably during the morning and noon in working days. This difference disappears in the weekends and therefore we associate it with the industrial and commercial activity. The electricity demand has a second peak in the evening, after working hours. Traffic workload evolves in a different way in the morning, increasing steadily until noon, then decreasing, and reaching its peak in the evening, together with the electricity demand peak.

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). If this assumption holds, then the ratios of electricity demand among locations are likely to be very similar to the ratios of web requests (this will also taking into account time difference between locations). Therefore, we allocate the web requests to each location following the ratios of electricity demand among the locations.

TABLE I
MAPE AND PRICES VS. BALANCED LOAD

	San Diego		Houston		New York	
GLB (%Load)	MAPE (%) & Avg. Price ($/MWh)					
0	3.04	47.90	2.75	43.94	3.06	70.22
5	6.77	49.31	3.46	45.51	6.39	70.80
10	8.15	49.83	7.24	47.23	7.63	71.01
15	10.74	50.81	10.54	48.76	8.63	71.19
20	14.28	52.15	14.81	50.77	10.73	71.56
MAPE/GLB	0.714		0.921		0.345	

Experiment Setup: We evaluate the prediction error of the utility. We change the demand corresponding to the allowed GLB workload between 0-15% of the total utility demand. We also extend the range up to 20% 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 [11] to allocate its allowed GLB workload optimally. We skip the details here.

The evaluation is carried out assuming that utilities use commonly adopted *neural networks* (NN)-based demand forecast algorithms [28] to predict their electricity demand[5]. Utilities use NNs as a black-box, which before being able to forecast the demand requires training with sample data, for which the right forecast is known. 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. 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).

Results and Observations: The results of how GLB affects retail prices are summarized in Table I. Each data center location has two associated columns. The first column shows the MAPE in the presence of varying GLB load percentage (increased at 5% resolution). The second column is the corresponding average retail prices according to Eq. 3. The last row shows the average MAPE per GLB load increment.

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 3.06%. A closer look into the prediction

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

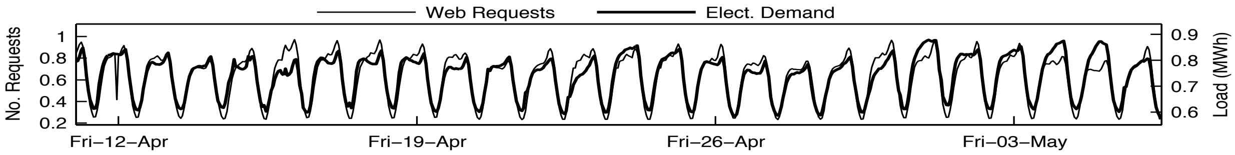
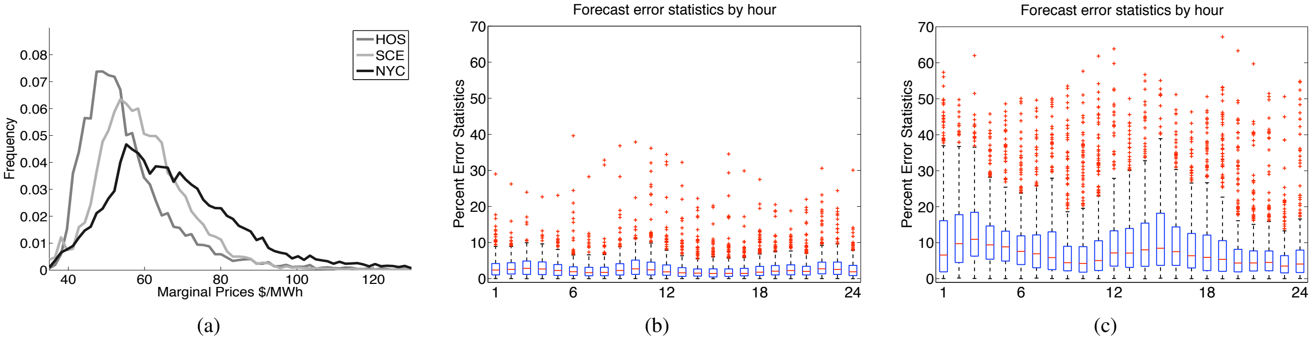


Fig. 2. Evolution of the (aggregated) electricity demand and web workload between April 12th and May 6th 2013.



(a) (b) (c)

Fig. 3. (a) MCP distribution for San Diego (SCE), Houston (HOS) and New York City (NYC) in 2009 - 2012; (b) MAPE for the demand prediction without GLB; (c) MAPE for the demand prediction with GLB at 10% (*i.e.*, the allowed demand variation caused by the CSP performing GLB is 10%).

accuracy of the NN algorithm 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 I, when the GLB load increases to 10%, 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.

Further details on the hourly MAPE distribution of the San Diego market for the cases of 0% and 10% GLB are shown in Figs. 3(b)-3(c). These box-plots show that the hourly MAPE distributions for the 10% GLB case are substantially worse than that of no GLB. For example, the 95-percentile of the MAPE (upper whisker of the box) is between 25-40%, as compared to 5-15% for the case of no GLB. A similar increment can be observed in the 25-percentile and 75-percentile results (upper/lower boundary of the box).

As MAPE gets worse upon GLB load percentage increment, utilities will increase the retail prices to ensure the same desirable profit as no GLB. This can be seen from Table I; **the retail price for San Diego on average increases by 0.7% for every 1% GLB load percentage increment.**

C. GLB Increases Electricity Bills for CSPs

To demonstrate how GLB ends up increasing electricity bills for the CSP, we continue our case study and compare the total electricity cost (sum of the three locations for the year 2012) for the cases where the CSP is able to move 0%, 5%, 10%, and 20% of the total local utility demand, which we denote as NOGLB, GLB@5, GLB@10, and GLB@20, respectively.

We compare the cost difference (in percentage) between the baseline case, NOGLB, and the rest. Results show that in the GLB@5 case **the CSP actually ends up paying a total bill 1% higher than not doing GLB at all**. In the GLB@10 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 GBL workload of its data centers. Finally, higher benefits could be achieved for large *allowed* GLB load. For the GLB@20 case, the GLB effect provides 9% cost reduction. However, this case requires the CSP to move up to 60% of its workload in each data center, which is beyond the *feasible* percentage in data centers nowadays (20-30% according to the estimate in [11]).

D. Summary

In essence, our analysis and case study show that GLB actually introduces trading inefficiency in today's supply chain, which leads to increased retail prices and higher electricity bills for CSPs. As such, a satisfactory GLB solution should avoid economic inefficiency, while allowing CSPs to exploit the advantage of GLB. The mechanism that we present in the next section allows the CSP to make a more optimal use of the electricity supply chain(s). We show that it can reduce the electricity cost for the CSP without causing economic loss to the utilities.

IV. A Broker-Assisted GLB Solution

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. By doing GLB and electricity procurement jointly, CSPs can eliminate the trading inefficiency between utilities and CSPs. Consequently, GLB no longer causes economic losses to utilities, and CSPs can still exploit their presence in multiple geo-locations to achieve desirable electricity cost reduction. In particular, this approach creates a "geo-distributed supply chain" illustrated in Fig. 1(c)), in which a CSP can first bid in different spot markets, *i.e.*, *balance its bids*, and then the CSP balance its load across geo-distributed data centers according to the obtained electricity supply, *i.e.*, *balance its bits*.

We remark that far from being a speculative scenario, the technique that we present in this section could be already used by some large CSPs like Google, which can buy and sell electricity directly from/to the spot markets since 2010 [15]. Meanwhile, utility companies (or other entities) may offer brokering services to CSP clients and avoid the GLB-induced economic loss.

A. 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 notations in the following. 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 [29].

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[6], then the electricity demand for serving $\sum_{i=1}^{m} u_{ij}a_{ij}$ amount of requests is simply $\gamma \cdot \sum_{i=1}^{m} u_{ij}a_{ij}$.

There are one day-ahead and one real-time market at data center location j and we use (b_j, q_j) as the bids that the CSP places in the day-ahead market auctions through brokers; here b_j represents the bidding price and q_j represents the bidding electricity quantity. The MCP for the day-ahead market and the real-time market are w_b^j and w_u^j, respectively, both of which are modeled as random variables with probability distribution function f_b^j and f_u^j, respectively. According to our discussion in Secs. II-A and III-A, the actual price that the CSP obtains for serving data center at location j, denoted as v_j, is a function of w_b^j, w_u^j, and b_j:

$$v_j = \begin{cases} w_b^j, & \text{if } b_j \geq w_b^j; \\ w_u^j, & \text{otherwise.} \end{cases} \tag{4}$$

That is, in case the bidding price b_j is higher than the MCP w_b^j, then the price is v_j; otherwise, the CSP has to purchase the electricity from the real-time market (by placing a very high bid so that it always gets the urgent supply at MCP w_u^j).

With the above notations, we can formulate the joint GLB and electricity procurement problem that the CSP needs to solve as follows:

$$\min \sum_{j=1}^{n} \mathbb{E}\left[v_j\right] \cdot \sum_{i=1}^{m} u_{ij}a_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}u_{ij} \tag{5}$$

$$\text{s.t.} \sum_{j=1}^{n} u_{ij}a_{ij} \geq U_i,\ 1 \leq i \leq m, \tag{6}$$

$$u_{ii} \geq \alpha \cdot U_i, 1 \leq i \leq m, \tag{7}$$

$$\sum_{i=1}^{m} u_{ij}a_{ij} \leq \min\left\{C_j, \frac{1}{\gamma}q_j\right\},\ 1 \leq j \leq n, \tag{8}$$

$$\mathbb{E}\left[v_j\right] = \int_0^{b_j} x \cdot f_b^j(x)dx + \mathbb{E}\left[w_u^j\right]\int_{b_j}^{\infty} f_b^j(x)dx, \tag{9}$$

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

In the above problem, the objective in Eq. 5 represents the total expected cost of energy procurement and network load balancing cost. The constraints in Eq. 6 say that the demand at every location i must be served. The constraints in Eq. 7 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. 8 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 q_j. This set of constraints and the objective function capture the coupling of energy procurement and GLB in minimizing the overall cost for the CSP. Eq. 9 is the closed-form formula of $\mathbb{E}\left[v_j\right]$ expressed in b_j, f_b^j, and f_u^j (since $\mathbb{E}\left[w_u^j\right] = \int_0^{\infty} x \cdot f_u^j(x)dx$).

Remark: If (b_j, q_j) $(1 \leq j \leq n)$ are given, then $\mathbb{E}\left[v_j\right]$ and $\min\left\{C_j, \frac{1}{\gamma}q_j\right\}$ are fixed and the above problem in Eqs. 5-9 reduces to the standard GLB formulation in [11], for which many efficient algorithms have been proposed in the literature.

Thus to solve the joint GLB and electricity procurement problem optimally, we use our proposed optimal-bidding algorithm as shown in Fig. 4 to obtain the optimal bids, denoted as $\left(b_j^*, q_j^*\right)$ $(1 \leq j \leq n)$, and then apply the known algorithms to solve the remaining GLB problem given $\left(b_j^*, q_j^*\right)$ $(1 \leq j \leq n)$.

B. An Optimal-Bidding Algorithm

A simple strategy to obtain a set of bids is to minimize $\mathbb{E}\left[v_j\right]$ and get the corresponding b_j (recall that it is a function

[6] For example, as reported by Google [11], each search consumes 0.28 Watts-hour electricity in its data centers.

```
OPTIMAL-BIDDING ALGORITHM
$\pi \rightarrow \hat{b}$, $0 \rightarrow Z$
for $k \in \{1, \dots, n\}$ do
    BID_PRICE(k, $\hat{b}$)$\rightarrow \mathbf{b}$
    BID_Q($k, Z, \mathbf{b}$)$\rightarrow q_k$
    AUCTION($k$)$\rightarrow (v_k, q_k)$
    $Z + q_k \rightarrow Z$
end for
/* Compute optimum GLB */
Solve Eq. 5 with prices $v_k$ and capacities $q_k$

BID_PRICE ($k, \hat{b}$)
$\min_l \left\{ \mathbb{E}[v_l](\hat{b}) \right\} \rightarrow s^*$
for $l \in \{k, \dots, n\} \backslash s^*$ do
    $\arg_b \left\{ \mathbb{E}[v_l](b) = \mathbb{E}[v_{s^*}](\hat{b}) \right\} \rightarrow b_l$
end for
return $(b_k, \dots, b_n) \rightarrow \mathbf{b}$

BID_Q ($k, Z, \mathbf{b}$)
$\min_l \{\mathbb{E}\{C_l(b_l, u_l)\}'|_{u_l=0}\} \rightarrow \nu$ s.t. $u_l = \sum_i u_{il}$
/* Solve the objective function for location $\nu$ */
$\arg_q \min \{\mathbb{E}\{C_\nu(b_\nu, q)\}\} \rightarrow q^*_\nu$
$\arg_{q_k} \{\mathbb{E}\{C_k(b_k, q_k)\}' = \mathbb{E}\{C_\nu(b_\nu, q^*_\nu)\}'\} \rightarrow q_k$
return $q_k$
```

Fig. 4. Optimal-Bidding Algorithm

of b_j) independently across locations $1 \leq j \leq n$. However, this naïve approach does not give the optimal b^*_j in general, since it does not take into account the ability that the CSP can move workload away from data centers at expensive locations. To see this, consider a simple scenario with $n = 2$ locations and the MCP of the first market is statistically much higher than that of the second market. In this case, it is more cost effective to only bid into the second market and serve all workload (assuming $\alpha = 0$ in Eq. 7) at the second location.

In our proposed optimal-bidding algorithm, we compute b^*_j jointly and sequently. To compute the optimum bid price for the next auction, the routine BID_PRICE takes into account the probability of winning in all the auctions in the sequence that has not been executed. We assume the CSP knows the marginal profit π per MWh consumed in the data center. For a single market bidding, the optimum bid price $\hat{b}$ is equal to the marginal profit π [17]. BID_PRICE computes each individual bid price b_j, so that the expected price value is the same for all auctions, *i.e.*, $\mathbb{E}[v_j](b^*_j) = \mathbb{E}[v_{s^*}](\hat{b})$. If not value satisfies the condition, then bid price is set to the marginal profit π.

The bid quantity is computed using BID_Q, which computes the optimum values u_{ij} in Eq. 5 for $q_j = C_j$ and the computed bid prices b_j. Similarly to BID_PRICE, the algorithm computes the bid values $q^*_j, \forall j$ that make the first derivative of the expected cost, denoted as $\mathbb{E}\{C_j(b_j, q_j)'\}$, equal in all the auctions. In our case the derivative $\mathbb{E}\{C_j(b_j, q_j)'\} = \mathbb{E}[v_j] + (\sum_i z_{ij}(\partial u_{ij}/\partial q_j))$. If no value satisfies the condition on the derivative, then the bid quantity is set to zero.

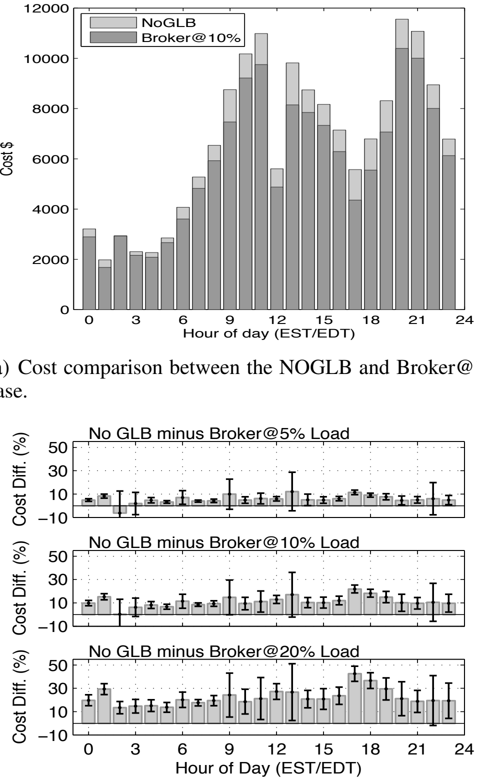
Afterwards, the auction is executed. The CSP bids at values (b^*_j, q^*_j). The algorithm is updated with the outcome of the auction. If the bid (b^*_j, q^*_j) is a winning bid, then the electricity $q_j = q^*_j$ is granted at day-ahead MCP prices w_j, otherwise at real-time MCP w^j_u. The algorithm continues for the remaining auctions. Once all auctions are executed, then the algorithm solves the problem in Eqs. 5-9 as the conventional GLB optimization, using the values $v_k, q_k, \forall k$.

The optimality of this algorithm depends heavily on the characteristics of the probability distribution of the markets and the network and capacity constraints, which are specific to each setting. The existing literature on optimal bidding in multiple auctions lacks of general techniques for arbitrary constraints [30], [31] and heuristics are used [32]. We provide an optimality theorem for our formulation under certain relaxed conditions.

Theorem 1: Assume the objective function in Eq. 5 is convex in b_j and u_{ij}, and C_j is sufficiently large so that it is not active in the constraints in Eq. 8. Then, the algorithm in Fig. 4 computes a set of optimal bids $\left(b^*_j, q^*_j\right)$ $(1 \leq j \leq n)$.

Proof: See the Appendix section. ■

Consequently, the cost minimization problem in Eqs. 5-9 can be solved by applying known GLB algorithms to solve the remaining GLB problem given $\left(b^*_j, q^*_j\right)$ $(1 \leq j \leq n)$.



(a) Cost comparison between the NOGLB and Broker@10 case.

(b) Hourly cost difference between NOGLB and the three *Broker* cases, @5%, @10% and @20% GLB.

Fig. 5. CSP cost for each hour of the day as 5(a) average cost and 5(b) differential w.r.t. the NOGLB case.

C. Performance Evaluation

We evaluate the performance of the algorithm in Fig. 4 in the scenario with three data-centers and dataset from Sec. III. We have implemented the algorithm and a simulator of sequential auctions, which we feed with our dataset.

We use the CSP's total electricity cost as the metric. We compute this cost in four cases using different percentages of allowed GLB load, *i.e.*, 0% (*NOGLB* case), 5%, 10% and 20% (*Broker@5*, *Broker@10* and *Broker@20* cases). For the Broker@5, @10 and @20 cases, attending distance constraints, the CSP runs the bidding algorithm for two markets, *i.e.*, $n = 2$, for San Diego(-Houston) and New York(-Houston) data centers load and $n = 3$ for Houston. Meanwhile for the NOGLB case, the prices and the cost are obtained in the same way as described in the previous section for fair comparison.

We first compute the average cost per hour of the CSP. Results for the NOGLB and Broker@10 ($\sim$30% of the data center workload) cases are displayed in Fig. 5(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 lunchtime and commuting after work and the peak hour around 8-9 pm.

The darker gray portion of the bars are the average cost for our broker-assisted solution. The fair gray portion of the bars represent the extra cost in average that the CSP pays in the NOGLB case, which is noticeable in most of the hours. We also study how the percentage of allowed GLB load increases the cost reduction. From top to bottom, Fig. 5(b) shows the Broker@5, @10 and @20 cases, respectively. According to the results displayed (more positive is better), the achieved cost reduction increases with the amount of balanced load. Hourly reductions between 5-10% can be obtained by moving 5% of the data center electricity demand ($\sim$15% of the CSP workload) and they can be increased further, 10-20%, if $\sim$30% of the CSP workload is geo-load balanced and 10-40% for the unrealistic Broker@20. Computing the absolute cost for a year for the Broker@5 and @10, **our broker-assisted solution provides 6-12% annual savings**, compared to NOGLB.

V. 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.* [11]. Subsequent publications analyzed the technical feasibility and assess the possibilities of GLB [9], [10], [12], [33], [34]. 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 [35]. 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 contrast to the utilities, the CSP is able to bid in markets in different locations and constraints, providing new study case for the existing literature on strategic bidding [31], [32], [17].

VI. 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 – 6-12% less than not doing GLB, for our case study based on real-world traces.

References

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

[2] "Spain energy consumption," http://www.nationmaster.com/country/sp-spain/ene-energy.

[3] L. Barroso and U. Holzle, "The case for energy-proportional computing," *IEEE Computer*, vol. 40, no. 12, pp. 33–37, 2007.

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

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

[6] 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.

[7] 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*, vol. 36, no. 1, 2008, pp. 48–59.

[8] 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, M. Zelkowitz(ed.), vol. 82, pp. 47-111*, 2011.

[9] Z. Liu, M. Lin, A. Wierman, S. Low, and L. Andrew, "Greening geographical load balancing," in *Proc. ACM SIGMETRICS*, 2011, pp. 233–244.

[10] P. Wendell, J. Jiang, M. Freedman, and J. Rexford, "Donar: decentralized server selection for cloud services," in *Proc. ACM SIGCOMM*, vol. 40, no. 4, 2010, pp. 231–242.

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

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

[13] "How clean is your cloud?" Greenpeace Climate, Tech. Rep., 2012. [Online]. Available: http://www.greenpeace.org/international/en/publications/Campaign-reports/Climate-Reports/How-Clean-is-Your-Cloud/

[14] "2011 Oregon Utility Statistics." [Online]. Available: http://www.puc.state.or.us/docs/statbook2011.pdf

[15] "Google energy wiki," http://en.wikipedia.org/wiki/Google_Energy.

[16] "2012 state of the markets report," Federal Energy Regulatory Commission, Tech. Rep., 2012. [Online]. Available: http://www.ferc.gov/market-oversight/reports-analyses/st-mkt-ovr/2012-som-final.pdf

[17] M. Liu and F. Wu, "Risk management in a competitive electricity market," *International Journal of Electrical Power & Energy Systems*, vol. 29, no. 9, pp. 690–697, 2007.

[18] P. Joskow, "California's electricity crisis," *Oxford Review of Economic Policy*, vol. 17, no. 3, pp. 365–388, 2001.

[19] "Duke energy annual report 2011," http://www.duke-energy.com/pdfs/DukeEnergy_2011_AR-10k.pdf.

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

[21] "Akamai Internet Observatory website." [Online]. Available: http://www.akamai.com/html/technology/dataviz3.html

[22] "Facebook's new 'cloud'," ECONorthWest, Tech. Rep., 2011. [Online]. Available: http://www.econw.com/

[23] J. M. Camacho, M. Chen, and D. M. Chiu, "Balance your bids before your bits: The economics of geographic load-balancing," Tech. Rep., 2013. [Online]. Available: http://www.ie.cuhk.edu.hk/~mhchen/papers/GLB.tr.pdf

[24] "Caiso archive," available at http://www.caiso.com.

[25] "Ercot archive," available at http://www.ercot.com.

[26] "Nyiso archive," available at http://www.nyiso.com.

[27] "Weatherunderground," http://www.wunderground.com.

[28] G. K. Tso and K. K. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, 2007.

[29] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao, "Energy-aware server provisioning and load dispatching for connection-intensive internet services," in *Proc. USENIX NSDI*, 2008.

[30] E. Gerding, R. Dash, D. Yuen, and N. Jennings, "Optimal bidding strategies for simultaneous vickrey auctions with perfect substitutes," in *Proc. 8th Int. Workshop on Game Theoretic and Decision Theoretic Agents, Hakodate, Japan*, 2006, pp. 10–17.

[31] 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*, vol. 1, no. 4, pp. 443–463, 2005.

[32] 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*, vol. 27, no. 3, pp. 1204 –1213, aug. 2012.

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

[34] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, "It's not easy being green," in *Proceedings of the ACM SIGCOMM 2012*, 2012, pp. 211–222.

[35] 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*, 2013, pp. 414–421.

Appendix

In order to prove the optimality of our algorithm, we first characterize the derivative of expected cost function, *i.e.*,

$$\mathbb{E}\{C\} = \sum_{j=1}^{n} \mathbb{E}\{C_j(v_j, q_j)\} \tag{10}$$

$$= \sum_{j=1}^{n} \mathbb{E}\{v_j\} \sum_{i=1}^{m} u_{ij} a_{ij} + \sum_{i=1}^{m} \sum_{j=1}^{n} z_{ij} u_{ij}, \tag{11}$$

and the constraints to the optimization problem by deriving the KKT conditions.

$$U_i - \sum_{j}^{n} u_{ij} a_{ij} \leq 0 \tag{12}$$

$$\alpha \cdot U_i - u_{ii} \leq 0 \tag{13}$$

$$\sum_{i} u_{ij} a_{ij} - \min\{C_j, \frac{1}{\gamma} q_j\} \leq 0 \tag{14}$$

$$-\lambda_i \leq 0 \tag{15}$$

$$\lambda_0(\alpha \cdot U_i - u_{ii}) = 0 \tag{16}$$

$$\lambda_1(U_i - \sum_{j}^{n} u_{ij} a_{ij}) = 0 \tag{17}$$

$$\lambda_2(\sum_{i} u_{ij} a_{ij} - \min\{C_j, \frac{1}{\gamma} q_j\}) = 0 \tag{18}$$

$$\frac{1}{\gamma}\left[a_{ij}\mathbb{E}\{v_j\} + z_{il}\right] + \lambda_1(-a_{ij}) + \lambda_2(a_{ij}) = 0 \tag{19}$$

From this equations it is easy to verify that setting the Lagrange multipliers $\lambda_i = 0$, the conditions with terms in the multipliers hold. As a consequence, we can also state that there is no optimality gap and the optimum of the unconstrained objective function matches the optimum of the problem as long as the rest of the KKT conditions hold.

Define the quantity $q_j = \gamma(\sum_i u_{ij} a_{ij}),\ \forall j$, such as the electricity purchased in location j at (expected) price $\mathbb{E}\{v_j\}$, then **the optimal allocation (according to the optimality condition from [31])** q_j^* **is such that makes the marginal expected cost (*i.e.*, first derivative of the expected cost) equal in all the markets**.

The intuition behind this condition is that we allocate quantities following a water-filling principle, allocating first as much electricity as we can to locations where the expected marginal cost is lower and once we cannot allocate more, then we continue allocating the remaining electricity (if any) to the second *cheapest* location and so on.

If we assume for the moment that network costs are negligible, data center capacities arbitrarily large and the bidder is not updated with the results of previous auctions, then the expected cost in each location is the product of the (expected marginal) price and the quantity of electricity purchased to power the data center. In this simplified case, we can express the first derivative like,

$$\frac{\partial \mathbb{E}\{C_j(v_j, q_j)\}}{\partial q_j} = \frac{1}{\gamma}\mathbb{E}\{v_j\} \tag{20}$$

Since $\mathbb{E}\{v_j\}$ is by definition a function of the bid prices b_j, the algorithm (see BID_PRICE routine) makes the expected prices minimum and equal in all the markets, *i.e.*, $\mathbb{E}\{v_j\} = \mathbb{E}\{v_i\}\ \forall i, j$. Note that, by making this expected prices the same in every market, we are fulfilling the optimality condition in [**?**] for separated bidding. In that case, the optimum bidding strategy is to split the purchase of the total demand D evenly among the auctions so that the strategy is $(b_j^*, q_j^* = D/n)$.

Provided that, we are doing the bidding jointly and the bidder knows the outcome of the previous auctions before bidding in the next one, we can re-define the probability density f_b^j of winning D in j-th auction as depending on the history of previous auctions, *i.e.* $f_b^j(v_j, b_j|q_{1,j-1})$, so that the expected marginal cost fulfills the following *newsvendor* equation (see [?] for a detail explanation of how to obtain this expression),

$$\frac{\partial\mathbb{E}\left\{C_j(v_j,q_j|i_{1,j-1}\right\}}{\partial q_j} = -(p_{j-1}+h)F_{j-1}(q_0+q_j)+p_{j-1} \tag{21}$$

,

where $h = 0$ is the buy back price at which over-supply is compensated (which in our case is zero because we assume perfect demand estimation) , p is the price at which we compensate the under-supply, which in our case is the expected price in the following day-ahead market, *i.e.*, $p_{j-1} = \mathbb{E}\{v_{j-1}\}$; and $F_{j-1}(q)$ is the cumulative distribution function of the demand, *i.e.*, the probability of having obtained a quantity q after auction j has been executed. Therefore (removing the factor in γ) we have that,

$$\frac{\partial\mathbb{E}\left\{C_j(v_j,q_j|i_{1,j-1}\right\}}{\partial q_j} = \mathbb{E}\left\{v_j|i_{1,j-1}\right\}. \tag{22}$$

Replacing in Eq. 21, we obtain

$$\mathbb{E}\left\{v_j|i_{1,j-1}\right\} = -p_{j-1}\cdot F_{j-1}(q_0+q_j)+p_{j-1}, \tag{23}$$

where

$$\mathbb{E}\left\{v_j|i_{(i,j-1)}\right\} = \left\{\begin{array}{ll} \mathbb{E}\left\{v_j\right\} & (q_0+q_j) < \sum_k q_k \\ 0 & \hfill otherwise \end{array}\right. \tag{24}$$

.

The previous equation means that the expect cost is only higher than zero the if the total required electricity has not been purchased yet in the previous auctions to auction j. Note that if the expected cost is zero, then Eq. 23 we have that

$$0 = -\mathbb{E}\left\{v_{j-1}\right\}F_{j-1}(q_0+q_j)+\mathbb{E}\left\{v_{j-1}\right\}. \tag{25}$$

Operating the expression we get to the following condition,

$$F_{j-1}(q_0+q_j) = 1 \quad \text{iff} \quad q_0 = \sum_k q_k \;\text{ and } q_j^* = 0. \tag{26}$$

This result can be interpreted as follows. The cumulative probability equals to 1 if and only if the entire desired probability has been purchased before the auction clears.

On the other hand, if the cumulative probability is equal to 0, then it means that no electricity has been purchased before, such that,

$$\mathbb{E}\left\{v_j\right\} = -\mathbb{E}\left\{v_j\right\}F_{j-1}(q_0+q_j)+\mathbb{E}\left\{v_j\right\} \tag{27}$$

and operating,

$$F_{j-1}(q_0+q_j) = 0. \tag{28}$$

Therefore, the optimum bidding quantity is,

$$q_j^* = q_0 + F_{j-1}^{-1}(0)\Big|_{=0} = q_0 \tag{29}$$

and moreover, $q_1^* = \sum_k q_k$. Hence, for the case of no network costs, since the marginal prices are constant in the amount, we get that the optimal bidding strategy is to bid the desired amount in all the markets in the sequence until the bid is a winning bid in one of them.

However, if we include the network costs, the expected marginal cost is not a fixed expected marginal price for any amount. Here is the expression of the derivative in q_j,

$$\frac{\partial\mathbb{E}\left\{C_j(v_j,q_j)\right\}}{\partial q_j} = \frac{1}{\gamma}\left[\mathbb{E}\left\{v_j\right\}+\sum_{l=1}^{n}\sum_{i=1}^{m} z_{il}\left(\frac{\partial u_{il}}{\partial q_j}\right)\right]. \tag{30}$$

The interesting aspect of this expression is that the marginal cost increases as we increase the amount of allocated demand u_{ij} to one location. To understand this, imagine that network costs increase linearly with the distance between users and a particular data center, *e.g.*, j. If we start with a small part of the total service demand allocated to data center j, the demand will be conformed with the requests from the closest users. As we increase more and more the demand, the demand comprises users from further distances, therefore with higher network costs.

This increases the expected marginal prices as we allocate more demand to the data center and it makes more economically attractive to allocate demand to other data centers. As a consequence, even if we are still bidding in sequential auctions, the optimal bid quantity cannot be the total service demand D for all the auctions.

In order to compute the optimum bid price we follow the sub-routine BID_Q. The first step (line 18), the routine computes the market with the lowest marginal cost if no electricity is purchased, that is, $\mathbb{E}\{v_j\}$. Since we have make that $\mathbb{E}\{v_j\} = \mathbb{E}\{v_i\}$ for all the auctions, then we can choose any, *e.g.*, k. Define $(j-1)$-th as the last auction before the the current one j-th. In line 19 the algorithm computes the optimum bid quantity, in our case q_k^* solving the linear program in Eq. 5 using the electricity prices and quantities obtained in previous auctions. Then, in order to compute q_j^* it makes equal the expected marginal costs for the k-th and j-th auctions. For that, we can use again Eq. 21. Denote the electricity obtained in the previous auctions as $\hat{q}_{j-1}$, then we obtain the following expression,

$$\begin{array}{l}\left[\mathbb{E}\left\{v_j|i_{1,j-1}\right\}+\sum_{l=1}^{n}\sum_{i=1}^{m} z_{il}\left(\dfrac{\partial u_{il}}{\partial q_j}\right)\right] = \\ = -p_{j-1}\cdot F_{j-1}(\hat{q}_{j-1}+q_j)+p_{j-1}\end{array} \tag{31}$$

.

If the electricity has been purchased before the j-th auction, the left term is null and it is easy to check that $q_j^* = 0$. Otherwise, the optimum bidding amount is given by,

$$q_j^* = F_{j-1}^{-1}\left[\hat{p}\right] - \hat{q}_{j-1} \tag{32}$$

$$\hat{p} = -\frac{1}{p_{j-1}}\left(p_{j-1} - \left[\mathbb{E}\left\{v_j|i_{1,j-1}\right\} - \sum_{l=1}^{n}\sum_{i=1}^{m} z_{il}\left(\frac{\partial u_{il}}{\partial q_j}\right)\right]\right), \tag{33}$$

and since $p_{j-1} = \mathbb{E}\{v_{j-1}\}$ we have that,

$$q_j^* = F_{j-1}^{-1}\left[\frac{1}{\mathbb{E}\left\{v_j\right\}}\left(\sum_{l=1}^{n}\sum_{i=1}^{m} z_{il}\left(\frac{\partial u_{il}}{\partial q_j}\right)\right)\right] - \hat{q}_{j-1}, \quad (34)$$

since this amount can be negative, we re-define the expression of the optimum bid price as,

$$\hat{q}_j^* = \left\{ \begin{array}{ll} q_j^* & q_j^* > 0 \\ 0 & q_j^* \leq 0 \end{array} \right. \quad (35)$$

To complete the proof we need to show that this result verifies the condition in line 21. For that, recall that the probability of winning electricity from a market is a scalar value σ determined by the bid price and the probability distribution function of the market, *i.e.*, f_j^b. Therefore, the inverse of the cumulative function, $F^{-1}(\cdot)$, is a line with slope $1/\sigma$. Therefore,

$$q_j^* = q_k^* \quad (36)$$

$$\frac{1}{\mathbb{E}\left\{v_j\right\}}\left(\sum_{l=1}^{n}\sum_{i=1}^{m} z_{il}\left(\frac{\partial u_{il}}{\partial q_j}\right)\right) = \frac{1}{\mathbb{E}\left\{v_k\right\}}\left(\sum_{l=1}^{n}\sum_{i=1}^{m} z_{il}\left(\frac{\partial u_{il}}{\partial q_k}\right)\right) \quad (37)$$

Since the values $\mathbb{E}\left\{v_j\right\} = \mathbb{E}\left\{v_k\right\}$ then,

$$\sum_{l=1}^{n}\sum_{i=1}^{m} z_{il}\left(\frac{\partial u_{il}}{\partial q_j} - \frac{\partial u_{il}}{\partial q_k}\right) = 0 \quad (38)$$

This equation is interpreted as follows. The increment (decrement) of demand units allocated to a particular location l from a location i per increment (decrement) in the electricity purchase at location j must be the same as for the purchased electricity in the location k. This leads to the water-filling process that we referred in the description of the algorithm. For the unconstrained case, since the optimal solution to Eq. 5 will allocate more units to locations where the electricity is cheaper, this condition holds as the demand can be freely moved from one location to another. For the constrained case, from the KKT conditions, the solution provided by the algorithm is optimal as long as there is enough (purchased) capacity to serve all the demand (which is one of the assumptions in the theorem). Otherwise, the KKT conditions in Eqs. 12-14 may be violated.