Greening Cloud Service Pricing

Debdeep Paula\*, Wen-De Zhonga, Sanjay K. Boseb

aSchool of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

bDepartment of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati 781039, India

|  |  |  |
| --- | --- | --- |
| A R T I C L E I N F O  Article history:  Received 00 December 00  Received in revised form 00 January 00  Accepted 00 February 00  Keywords: |  | A B S T R A C T |

1. Introduction

Over the past few decades there has been a remarkable progress in various aspects of Internet technology as a whole and a as an obvious result these days various sophisticated services are provided online. To cope up with that unprecedented growth of technologies the data centers (which are the physical resources) behind the cloud computing technology are also going at an alarming pace. This is continuing with the advent of innovative technologies such as Internet of Things (IoT), Mobile Computing the data centers are expected to grow even at a higher rate, in near future. With this massive growth of Internet technology, the power consumed by the data centers, which are the primary infrastructures for cloud services is growing dramatically. As a result the cost of electricity constitutes a substantial proportion of the total operational cost. For example, Google officially revealed, their data centers located at various different parts of the globe cumulatively consumed $138M worth of electricity annually [[1](#_ENREF_1), [2](#_ENREF_2)] in 2011, which would have increased further by now. However, the more alarming fact is that the electricity consumption is growing at an exorbitantly high rate. In US alone, the growth in electricity demands across the ICT sector is ten times higher than the aggregated growth of the electricity demand [[3](#_ENREF_3)]. Also, The total annual electricity cost of servers and data centers in the United States is estimated at $7.4 billion [[4](#_ENREF_4)].

Given the importance of electricity cost in the business model of a Cloud Service Provider (CSP), quite surprisingly the pricing for the various cloud services all across Iaas, Paas and Saas mostly remains insensitive to various aspects which contributes to the power cost, incurred by the CSP. We believe, the pricing policy implemented by various different CSPs can be largely improved by addressing the real time price of electricity in the wholesale market and the renewable energy generation. The high level goal of this work is to provide some valuable insights on how a CSP can price the services provided to the consumers to optimize the revenue and profit, in a competitive market where multiple CSPs are competing among themselves and the consumes are price sensitive. Additionally, we aim to incentivize integration of renewable energy through this pricing strategy.

In this work we make the following contributions:

* We design an energy aware resource pricing scheme for multiple cloud service providers (CSP) addressing competition and derive the analytical expressions for the equilibrium load balancing condition under reasonable assumptions in a competitive market, where multiple CSPs operate and the consumers are rational and price sensitive.
* We construct a bilevel optimization problem, through which, at the slower timescale the service demands are distributed among multiple CSPs competing with each other.
* Once the service demands are distributed, then we propose an efficient algorithm to carry out a faster timescale optimization problem to schedule the deferrable jobs to maximize the renewable energy integration, based on Receding Horizon Control (RHC).
* We evaluate the proposed methodologies and algorithms with traces of workload collected from real server clusters, real life renewable traces in order to emulate the practical aspects of a CSP.

In this work we present a two stage optimization problem. One of them is solved in a slower timescale which addresses the distribution of service demand among the competing CSPs, based on a discrete time model. Since, in the wholesale market of electricity the price of electricity varies in one hour interval, we formulate and solve the slower timescale optimization problem once in an hour. The output of the slow timescale optimization problem are demand for services at different CSPs by considering electricity price, predicated renewable energy supply, power consumed by non-IT infrastructure. To address the competition among different CSPs we model the demand served by a CSP as a function of service quality offered by the CSP, price charged and the price charged by the competitor. By this we structured a fairly exhaustive model of a cloud service market, with multiple CSPs competing among themselves, with price sensitive rational consumers. In the fast timescale optimization we propose a Model Predictive Control (MPC) strategy to perform the scheduling of deferrable jobs to maximize the renewable energy integration into the grid. Since, the supply of renewable energy varies at a faster timescale, to address the variability, intermittency and randomness we proposed this two timescale bilevel optimization problem.

In our work we leverage some of the smart grid functionality to realize our objectives. It is well recognized that big data centers are excellent candidates to contribute towards environment sustainability. They are big consumers of electricity and a significant proportion of the workload is deferrable. As a matter of fact there is an excellent opportunity to implement *demand response* (DR) to shape the demand in line with supply. DR is a very efficient tool to integrate *random* renewable energy into the grid. The work in [[5](#_ENREF_5)] presented a nice summary of various DR strategies proposed in the literature and practiced across various different domains. The pricing mechanism is formulated to incentivize renewable energy integration into the grid. When one CSP has access to more renewable energy, it can offer a lower price to the customers and this would increase demand for that CSP, instead of deploying decremental reserves. A challenge is to select appropriate pricing policies. A small change in price may not be effective in incentivizing renewable energy integration whereas too high changes in price may result into system oscillation which would make the cloud service market reasonably unstable and inefficient. We take a two timescale approach. In the slower timescale we perform the load balancing and in the faster timescale we perform the job scheduling.

Moreover, most of the previous strategies can only address only colocated sources of renewable energy, most often owned by the data center operator. While in our work we formulate the model such that both onsite and offsite renewable energy sources can be incentivized. These days some big organizations owning data center rely on utility service providers for supplying the renewable energy. Typically they have a long term contract with the utility company to contribute towards environment sustainability and meet the (sometimes quite aggressive) carbon emission goals set by the various regulating bodies.

The rest of the paper is organized as follows. In Section 2, we present the related works and discuss typical state of the art practices and their recent improvements. In Section 2, we introduce the problem and present the model with a cloud service market with multiple CPSs and price sensitive consumers. In this Section we present the approach of load distribution at a slower timescale. Next, in Section 3, we present the scheduling problem which is applied in the faster timescale. In Section 5, we describe the setup for evaluation of the proposed approaches and discuss the results. Section 6, concludes the paper and presents some interesting future directions to address meaningful issues.

1. Related work

The powerful concept of cloud computing has started being realized through practical implementation since a few decades. During the entire phase of evolution both the concepts and actual implementation has gone through dramatic metamorphosis. An excellent discussion on the various aspects of cloud computing has been presented in [[6](#_ENREF_6)], which addresses both the technological aspects and business perspectives. A good source of reference touching upon various engineering, operational and business aspects of data centers is [[7](#_ENREF_7)]. The work in [[8](#_ENREF_8)] presents a fairly exhaustive survey on various resource management strategies proposed in the literature and practices prevalent in industry. However, the focus of our work is on the economics and pricing side of the cloud service markets. In the bilevel optimization problem, the slower timescale optimization problem address various aspects on the economic side of the cloud service market. There are various kinds of pricing strategies proposed in the literature and some of them are adopted by the industry. However, mostly static pricing strategies are implemented by most of the CSPs operating in the market today. A pioneer work suggesting an efficient autonomic pricing mechanism which dynamically adjusts the pricing parameters to discriminate the different service and application requirements of the different consumers. The most important advantage of customized, dynamic pricing is that, they can address the resource contention in a better way and thereby can improve on resource usage efficiency, which results into higher revenue of the CSP. Being a pioneer in cloud service market Amazon EC2 has introduced a *spot pricing* strategy where excess capacity (servers, storage capacity) are sold at a dynamic price to better leverage supply and demand [[9](#_ENREF_9)]. Although, this pricing strategy is quite straightforward and easy to implement. Nevertheless a recent work [[10](#_ENREF_10)] reported that spot pricing is quite inefficient to reflect the actual market demand by performing an exhaustive empirical analysis. They tackled the revenue management issue through a framework based on economics and formulated a stochastic optimization problem for revenue maximization with dynamic pricing. In general the theoretical work on service pricing within a cloud framework is quite rich. The work in [[11](#_ENREF_11)] presented an analytical framework focusing on the fairness and revenue tradeoffs that arise for a heterogeneous set of physical and virtual resources of a data center and the influence of various pricing strategies can have on the tradeoff between fairness and revenue maximization. They also characterized the implications of various different pricing strategies on different fairness metrics and derived the analytical bounds on the CSPs strategies to balance between fairness and revenue.

The work in [[12](#_ENREF_12)] presented synergetic framework where scheduler, charge back model and work in harmony to promote green and energy efficient computing systems. They also provided evidences on the feasibility of the proposed strategies by performing experimental evaluations. Another work along the same line [[13](#_ENREF_13)] presented a cost optimization framework, with collaboration between CSP and the utility provider, aiming to increase the renewable energy penetration. The underlying idea is to implement appropriate pricing as a control signal to address the renewable energy incorporation issue. However, in their work they have considered a single CSP. Additionally, they did not explicitly consider renewables, just reflect the renewable generation through price demand and supply difference. Moreover, they did not account for the cost for server switching (toggling from ON to OFF or vice versa).

The work [[14](#_ENREF_14)] considers the competitive market for cloud services serves as an important theoretical references for us. However, they primarily focus on *economics* of cloud service markets and we focus on formulating pricing strategies which makes the system greener. To realize that high level objective we leverage on some effective strategies prevalent in smart grid arena. Additionally, we merge the pricing (performed in slower timescale) with job scheduling (performed in faster timescale). We also augment the work in [[14](#_ENREF_14)] by considering a more realistic model which reflects the adverse effects of server switching. We formulate and solve convex optimization problems at different timescales which are computationally efficient.

The second level of the bilevel optimization problem consists of a fast timescale optimization problem which address the issue of renewable energy incorporation. The various different aspects of the power efficiency related issues in single or multiple data centers are addressed by many groups across the academia and industry. One of them is electricity cost reduction, which we also address in this paper. The pioneer work [[2](#_ENREF_2)] proposed to leverage spatial and temporal variation of electricity price in the wholesale market, to reduce the electricity cost incurred by a CSP owning multiple geographically distributed data centers. The two notable works [[15](#_ENREF_15), [16](#_ENREF_16)] merged the aim of maximizing the renewable energy integration with the cost reduction aspect and largely popularized the effectiveness of Geographical Load Balancing (GLB) in improving on energy efficiency of a CSP with multiple geographically distributed data centers. The aspect of renewable energy integration has nicely been addressed by [[17](#_ENREF_17)], by intelligent job scheduling and shaping the demand in line with generation of renewable energy. A review paper [[18](#_ENREF_18)], gives a fairly exhaustive overview of the present proposed and in practice approaches to incorporate renewable energy, to power both single and multiple geographically distributed data centers.

1. Problem setup-cloud market model with multiple CSPs

In this section we briefly introduce the cloud service market and explain the model we formulate to capture various aspects of the cloud service market.



Fig.1: Interaction between the three major entities in the cloud service markets

The market for cloud services are still in a nascent stage and as the whole market grow in size and scale then various other trading and pricing mechanisms are likely to come up [[19](#_ENREF_19)] to create an efficient market with healthy competition between the different CSPs and price sensitive rational consumers. We assume a very straight forward market with three entities, i) A set of regulatory bodies, ii) multiple CSPs competition among themselves to generate profit and iii) price sensitive rational consumers who generate the service demands in the market. Each of these three entities has different objectives and accountabilities. This is quite a general model. In reality the different services are categorized as, Infrastructure as a service (IaaS), Platform as a service (PaaS) and Software as a service (SaaS), and sometimes they have different business models [[20](#_ENREF_20)]. The objectives and accountabilities of the three above mentioned entities are presented in Fig.1. The market regulator is responsible for auditing the service quality provided to the consumers and ensuring a healthy competition between the CSPs operating in the region. They are also responsible for maintaining a stable and efficient where no CSP can become a monopoly or manipulate the price paid by the consumers by behaving strategically. On the other hand if the competition is too fierce then some of the CSPs would not be able to sustain in the market and the market will be unstable. However, the main objective of this paper is to provide some insightful information to the CSPs operating in market, about the optimal way to determine service prices to strike a balance between earning profit and attracting more consumers. The inherent trade-off is that the CSP should earn sufficient profit (Revenue - cost incurred) to sustain the business and at the same time they should not charge too much to the consumers which would result into drastic loss of market share (demand for services) and would eventually will get extinct. Additionally, the CSPs should provide acceptable level of Quality of Service (QoS) to the consumers. Although, the ultimate objective of a CSP is to earn profit but too aggressively pricing services would be detrimental to the business sustainability of the CSP. On the other hand we assume the consumers are price sensitive and would choose the CSP which would provide better service in less cost. For simplicity, we assume the price information of different CSPs operating in the market. We plan to consider the market with incomplete information, in our future work.

1. Distribution of service demand among the multiple CSPs: Slower Timescale Optimization Problem (*STOP)*

Table 1: The notations used in *STOP* and their meanings

|  |  |
| --- | --- |
| Notation | Meaning |
| *T* | Time index of *STOP.* |
| *Ω* | Time horizon of *STOP,* in the unit of *T.* |
|  | Service quality offered by CSP *i,* during timeslot *T*. |
|  | Maximum allowable delay for a job, during timeslot *T*. |
|  | Average delay experienced by a job serviced by CSP *i*, during timeslot *T*. |
|  | Aggregated service demand generated from the consumers at time instant *τ.* |
|  | Service demand (amount of jobs serviced) experienced by CSP *i* at time instant *τ.* |
|  | Average service demand (amount of jobs serviced) experienced by CSP *i* during timeslot *T*. |
|  | Average aggregated service demand generated from the consumers during timeslot *T*. |
|  | A function determined by CSP *i*, to reflect the sensitivity of demand with respect to service quality. |
|  | Sensitivity of demand with respect to the price experienced by CSP *i*. |
|  | Price determined by CSP *i*, to service an amount of demand, during timeslot *T*. |
|  | Sensitivity of demand experienced by CSP *i,* with respect to price charged by CSP *j*. |
|  | Utility of CSP *i*, for serving demand, during timeslot *T*. |
|  | Minimum price incurred by CSP *i,* for serving the required demand, during timeslot *T*. |
|  | Cumulative revenue earned by CSP *i* from timeslot 1 to *Ω*. |
|  | Cumulative profit earned by CSP *i* from timeslot 1 to *Ω*. |

Now we present the first level of the bilevel optimization problem. The output of this algorithm is the distribution of service demand among multiple CSPs operating in the cloud service market. The objective is to address the economic aspect of the market and to provide some insightful information which would help the CSPs to determine the price for the service, in this type of market. Throughout the rest of the paper we call this as, Slow Timescale Optimization Problem (STOP), since this problem is solved at the uniform interval of a longer time period. More specifically, at this level the CSP has the information on price of electricity, the prediction of renewable energy and the objective is to distribute the demand among all the CSPs operating in the market. This price can be set either by a centralized agency, such as the regulator or can be determined by exchange of information between the CSPs in a distributed manner. In this work we focus more on the interaction between the different factors to provide some insightful information about the structure of the optimal pricing schemes. For the sake of simplicity we rely on a centralized model for solving the optimization problem. However, the centralized system can certainly be extended to be solved in a distributed fashion by performing information exchange among the CSPs in an iterative manner.

The service qualityis usually measured in terms of the response time for the execution of the jobs. Usually, there is a maximum delay defined as per the Service Level Agreement (SLA) and we assume the QoS to be equal to the difference between the upper bound on the average response time and the actual average response time.

 (1)

In practice can be expectation or some percentile delay.

For example:

 (2)

Asincreases the QoS improves, so demand for services to CSP *i* increases. In other words the demand function (for a particular CSP) *x(.)* is an increasing function inLater in this section we describe about this function.

For the service demand we consider the average service demand requested by the clients, in aggregate and the amount of service demand served by CSP *i*, respectively, over the time period *T*. Mathematically, they are given by:

 and (3)

To model the demand function of a particular CSP we consider three aspects. They are namely, i) the QoS provided by the CSP, ii) the price it is charging to the consumers and iii) the price the competitor CSPs are charging to the consumers. This is in line with the model presented in [[14](#_ENREF_14)]. We perform a modification over that, which is to consider the *profit* made by the other competitive CSPs, i.e how much they are charging the consumers above the operational cost incurred by the CSP to provide the service as per SLA. This directly translates into the profit the CSP chooses to make. To determine the service price involves the inherent trade-off between profit and demand, in a competitive market with price sensitive rational consumers. If one CSP chooses to go for a higher profit then subsequently it would face reduction in demand. The main objective of this paper is to provide some insightful information to the CSPs to achieve a rational trade-off between profit margin and not to loose the market share. We control the sensitivity of the consumers with respect to other CSPs through the parameter  A higher value of implies the demand for CSP *i* increases sharply as the CSP *j* charges more to the consumers. This model is inspired by the model proposed in [[14](#_ENREF_14)]. However, the difference is that, they considered the demand is dependent on *price* the CSP *i*, is charging and the other CSPs are charging but we consider the demand is a function of *profit* CSP *i* is earning and the other CSPs are earning. One reason for the same is that, these days many IT giants are focusing more on operational cost reduction through technological improvements and this would give them a competitive advantage over the other players in the market. We model the demand function as per the following:

(4)

We capture the *utility* function in a novel way to capture the two most important factors affecting the business performance of a CSP, namely,

i) The revenue earned, which is a good indicator of financial performance in the short term.

ii) The amount of demand served by the CSP, which is a good indicator of the market share of the CSP in the competitive market and reflects comparatively less tangible parameters like, good will, market penetration etc, which translates into business sustainability.

It is to be noted that, in our model the pricealready takes the demand served into account. Nevertheless, to explicitly consider the market share we formulate the utility function as the product of *profit* earned and demand served by CSP *i*. Following that, the utility function of CSP *i*, is given by:

(5)

Since, this function is separable in price, we take the first order partial derivative with respect to

(6)

Now, equating (6) to zero, we obtain a relation between the optimum demand and the profit for CSP *i*.

The optimal price and optimal load are denoted by (\*) with the respective symbol.

 (7)

Equating (7) to (4) we obtain, an elegant expression for the optimum profit to be earned CSP *i*.

 (8)

We now, plug in a simple queueing theoretic formula to get some insight on the pricing strategy. Since, that is a simple and established way to related the number of servers to be deployed, the average workload (here service request) and the average response time. In case of data centers a popular queueing theoretic model is the M/M/k model, which is widely used in the literature [[21](#_ENREF_21), [22](#_ENREF_22)]. Now, we have the number of servers required with the average response time  as in (9) ;

 (9)

 (10)

We assume to be a simple linear function in and , with a multiplication constantwhich can be thought of as a knob to control the weightage of in the demand function in (4). Substituting the respective values we get:

 (11)

Combining (8) to (11) we have a fairly clean analytical expression for the optimum price to be charged by CSP *i*, as in (12).

 (12)

The overall revenue generated by CSP *i,* from T=1 to *Ω* is obtained by:

** (13)

The overall profit earned by CSP *i* is obtained by:

 (14)

Now we compute the operational cost incurred by the CSP to service the demands. Since the focus of this paper, is to make contribution on the energy efficiency side of the CSPs we consider the electricity cost. Recently, Apple has made an extremely ambitious commitment, to power the data centers (owned by them) by 100% renewable energy in the long run [[23](#_ENREF_23)]. As an instance, Google maintains its leadership in building data centers powered by renewable energy, as it significantly increased the renewable energy purchasing and investment both independently and through collaboration with various utility providers. Facebook also demonstrated the commitment to contribute towards environment sustainability, by locating a giant data center in Iowa driving the largest purchase of wind turbines in the world [[23](#_ENREF_23)]. The bottom line is that, the cost implication for integrating renewable energy varies dramatically from one to the other CSPs and almost impossible to model under one unified framework. Nevertheless, the model we present here is quite general and can incorporate every other operational costs as well. For the renewable energy, we extend this model to incorporate long term contract from the utility provider of renewable energy generator, which is typically the case for many IT companies adding substantial renewable energy to their energy portfolio. Without going into much administrative complexity we assume a simple incentive for integrating renewable energy. To control the value of the incentive we introduce a parameterassociated with CSP *i*. To address the special case with collocated renewable energy sources owned by the CSP with no administrative barrier can also be addressed by choosing equal to the spot market electricity price 

where  (15)

Now we present the optimization problem to be solved which we call *STOP*. We aim to minimize the total service cost paid by the consumers to service their aggregate demand. This is equivalent to maximizing the *value for money* from the perspective of the consumers, which can be perceived as the consumer’s utility function. We feel this would be the most prominent way to design a stable, transparent market with healthy competition between the CSPs with, where consumers receives fair services, maximizing their utility.

We construct the *STOP* as,

*Minimize* (16)

Subject to

**** (17)

 (18)

 (19)

The function (16) represents the objective function and (17) to (19) represent the constraints of the optimization problem. The computational complexity of the above problem depends on the structure of the functions involved. However, in our model, is easier to tackle since the objective function and the feasible set (defined by the intersection of the constraints) are convex in the objective function. So, there are reliable, efficient algorithms to solve the above problem numerically with very high degree of accuracy [[24](#_ENREF_24)]. Moreover, since this is a convex optimization problem the optimal solutions can be constructed by expanding as per KKT conditions [[24](#_ENREF_24)].

Now we consider a simple case to have some revelling insights into the structure of the optimal pricing policy to be adopted by the CSPs. We consider a market with two CSPs and a finite number of consumers generating service demand. Following the steps described in this Section, we derive the closed from expression for the optimum price to be charged by the CSPs, (20) to (22). We assume the demand does not exceed the sum of the individual capacities of the two CSPs.

 (20)

 (21)

More compactly,

  (22)

As mentioned in (22), the price set by CSP 1, depends on the vector consisting of

i) The cost incurred to service the demand

ii) The product of demand served and service quality

iii) The profit made by the other CSP

Among these three, the first and the third are quite obvious from general intuition. But, the third is more revelling and provides us valuable insights. The third term actually suggests the CSP can price the services offered according to the demand served (which translates into market share) and the QoS, which is the average response time. More importantly, the optimal price contains the product of these two parameters, which implies the CSP can either choose to serve more demand with a slightly worse QoS or less demand with a slightly better QoS. Usually, there is a defined standard for the QoS and CSPs compete for price. This strongly suggests, the CSP which would be able to operate at higher efficiency level and achieve internal cost reduction will always position itself better in the market, as compared to the other. At the same time, in this proposed model the cost incurred to service the demand request does not account for any profit component. The profit to the CSP is obtained from the second and third term in the expansion of (22). These profit comes from the amount of demand served by the CSP, the QoS provided to the consumers and the profit other CSPs are earning. The explicit presence of the term reflecting how much profit the CSPs are earning ensures the cloud service regulator (Fig.1) has distributed control over the market and the no CSP can charge an excessive amount to the consumers. Also, from the CSP’s point of view this pricing strategy prevents the CSP from charging too much to the consumers and loose the market share and eventually go extinct.

The first vector in (22) includes the model parameters defined by the market. In an emerging market, where the sufficient knowledge about the parameters is not available, these can be learned over a period of time through the optimum trade-off curve. In our numerical simulation we determine these parameters by trying out various combinations.

1. Energy aware scheduling for maximizing renewable integration: Faster Timescale Optimization Problem (*FTOP*)

In this section we describe the method of scheduling of deferrable jobs aiming to maximize the renewable energy integration into the grid. This problem is solved more than once within a single *STOP* window horizon. The aim of this section is to provide a mechanism to integrate the renewable energy integration issue, and other energy efficiency together is the pricing strategy *STOP*, to implement another layer of energy efficient optimization. In *STOP*, we did not consider the scope of time scheduling of the deferrable service request, which is quite common in case of actual data center workload. Few examples of this kind of deferrable jobs, data analytics on a large data set, scientific simulation, high quality medical image processing, and various other scientific computation based workload. The QoS of these types of jobs ia that, they have to be serviced before the user specified deadline, mentioned at the time of demand request. In literature, these types of deferrable jobs are referred as jobs with slacks [[25](#_ENREF_25)] and are quite well addressed in some previous works in the general area of *smart grid* [[26](#_ENREF_26), [27](#_ENREF_27)]. In this section we present an MPC based algorithm to

Schedule the deferrable jobs to maximize the consumption of renewable energy, which is the major high level objective of the smart grid community [[28](#_ENREF_28)]. A rich discussion about theoretical foundations of MPC and the implementation algorithms are found in [[29](#_ENREF_29)]. Application of MPC based job scheduling algorithms in other domains of smart grid are available in [[26](#_ENREF_26)]. In this Section we present the MPC based job scheduling algorithm to achieve higher renewable energy integration by leveraging the deferrable demand request.

Table 2: The notations used in *FTOP* and their meanings

|  |  |
| --- | --- |
| Notation | Meaning |
|  | Amount of brown energy used by CSP *i*, during timeslot *t*. |
|  | Number of servers switched from off state to on state. |
|  | Number of servers switched from on state to off state. |
|  | A constant to control the priority for switching cost (adverse effect) from off state to on state. |
|  | A constant to control the priority for switching cost (adverse effect) from on state to off state. |
| *t* | Time index of *FTOP.* |
| *ω* | The time horizon of *FTOP*. |
| *z* | Index of a deferrable task. |
| *Zt* | Active (unfinished) jobs which require service at time *t*. |
|  | Number of active servers during timeslot *t*, deployed by CSP *i*. |
|  | PUE of the data center owned by CSP *i,* during larger time period. |
|  | The renewable energy available to CSP *i*, during timeslot *t*, expressed in number of servers that can be powered by the renewable energy. |
|  | Jobs scheduled at timeslot *t,* for deferrable task *z*, performed by CSP *i*. |
|  | Total amount of jobs executed by CSP *i,* during timeslot *t*. |
|  | Total demand experienced by CSP *i*. |
|  | Total service requirement of deferrable task *z*. |
| *az* | Arrival time of deferrable task *z*. |
| *dz* | Deadline of deferrable task *z*. |

In *FTOP*, we address two crucial parameters namely, the amount of brown energy used and the number of times a server is switched from Off to On and vice versa. The commercial servers, typically deployed in today’s data centers are not at all energy proportional. In other words, they consume a significant amount of power, even when at idle condition [[30](#_ENREF_30)]. Given that, the most prominent approach to achieve energy efficiency in data centers is to turn off the extra servers during low demand period and to make the system power proportional. The main challenge to implement this is the fact, that the demand is uncertain and sometimes vary dramatically with time. As a result switching servers On and Off as per the demand, may incur very frequent server switching. The adverse effects of server switching are i) delay cost, ii) energy cost, iii) increased wear and tear, iv) perceived risk of reducing server lifetime. The notable previous works suggesting algorithms to achieve power proportionality, considering the switching costs in data center systems are [[31](#_ENREF_31), [32](#_ENREF_32)].

 (23)

 (24)

 (25)

The brown energy consumption is presented in (23). The number of servers switched from Off to On is obtained by (24), and from On to Off is obtained by (25), during timeslot *t*. In *FTOP*, we minimize a linear combination of (23) to (25) with a set a constraints to ensure QoS and operational feasibility. This model is not novel and was already implemented by [[31](#_ENREF_31)]. The novelty in this work is to integrate *FTOP*, with *STOP*, to construct a bilevel optimization problem to implement cloud service pricing policies which would promote energy efficiency in the cloud service market.

Minimize  (26)

Subject to:  (27)

 (28)

 (29)

 (30)

 (31)

The constraints in *FTOP*, are presented in (28) to (31). The total service demand served during *t*=*1* to *ω,* is equal to the total service demand arrived in the corresponding *STOP*, is ensured through (28). For all deferrable jobs *z*, the total amount of service provided during arrival and deadline has to be equal to the size of the job; i.e., all the deferrable jobs are to be serviced before it’s respective deadline is ensured through (29). Jobs scheduled in timeslot *t*, of the deferrable service request *z*, is non negative, is imposed through (30). All the deferrable job request arriving are served is ensured through (31).

1. Performance evaluation

In this section we describe the settings under which we have evaluated the methodologies and algorithms we proposed in this paper and discuss the results. We assume there are two CSPs operating in the market who are competing in the market as depicted in Fig. 1. The CSP-1 owns a data center in California and CSP-2 owns a data center in New York. The electricity price traces are collected from [[33](#_ENREF_33)]. We carry out the simulation for 7 days starting from 1st Oct 2010 to 8th Oct 2010. For the *STOP*, the time interval of solving the optimization problem is 1 hour, whereas for FTOP, it is 10 mins. In other words, in one hour timescale the demand distribution is performed and in 10 mins slot the energy efficient job scheduling is performed. The renewable traces are collected from [[34](#_ENREF_34)]. The overall service demand is collected from a commercial server cluster owned by Facebook, which was used in [[34](#_ENREF_34)]. The PUE was collected from a commercial data center owned by google, which was presented in [[34](#_ENREF_34)].

Benchmark revenue:

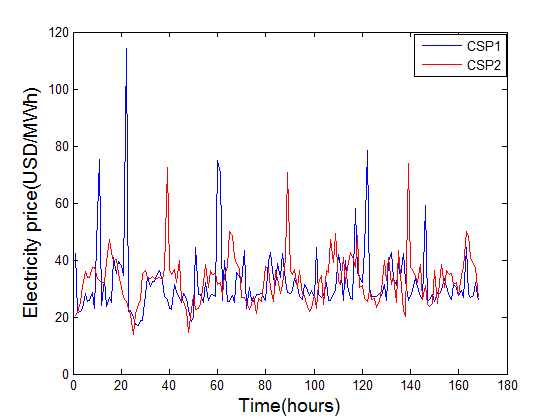
Let us define the two parameters,be the total revenue generated in the market (i.e., when aggregated across all the CSPs operating in the market) and**be the total profit earned aggregated across all the CSPs.

and ** (32)

The benchmark revenue for CSP *i*, is given by, and similarly, the benchmark profit for CSP *i*, is given by,

 (33)

The benchmark revenue and profit are the measure of what would have been the revenue and profit earned by CSP *i*, if there would have been uniform flat pricing, in the market. This pricing scheme is inefficient in terms of energy efficiency or promoting green cloud computing through pricing.



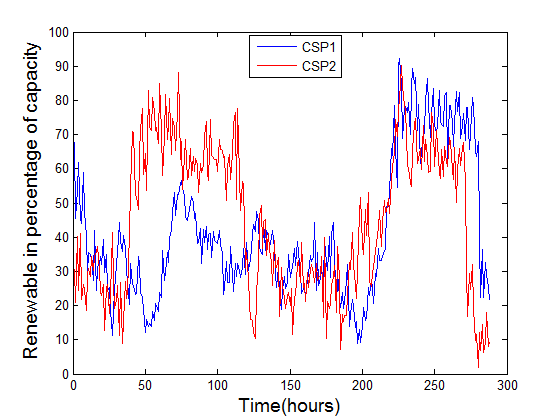


Fig.2: a) Electricity price in USD/MWatt hour used in the simulation, b) Renewable energy availability used in the simulation

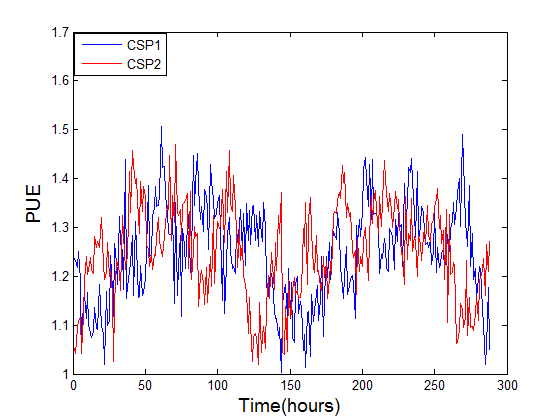
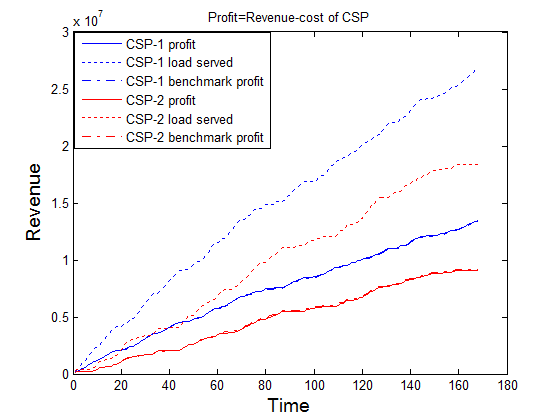


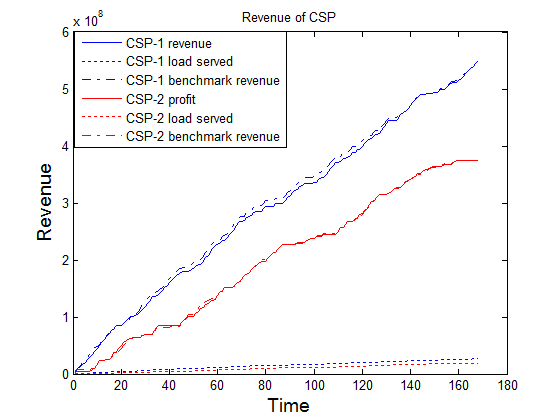
Fig.3: PUE of the two data centers owned by the CSPs.

However, this pricing scheme is quite straight forward, easy to implement and fair in terms of revenue and profit generation from the CSP point of view and fair in terms of service cost incurred by the consumers. So, we compare the economic efficiency of the proposed pricing scheme with the benchmark flat, uniform pricing scheme as in (32) and (33). The results are presented and discussed below.

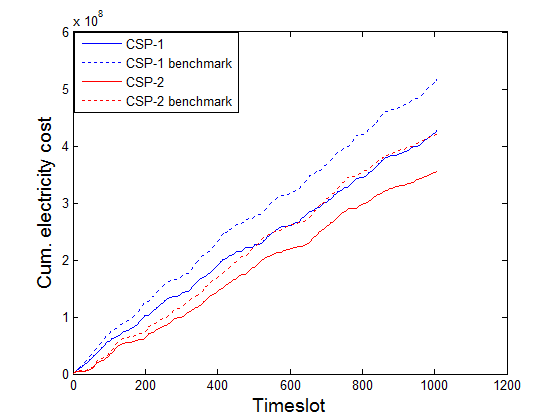
Results:

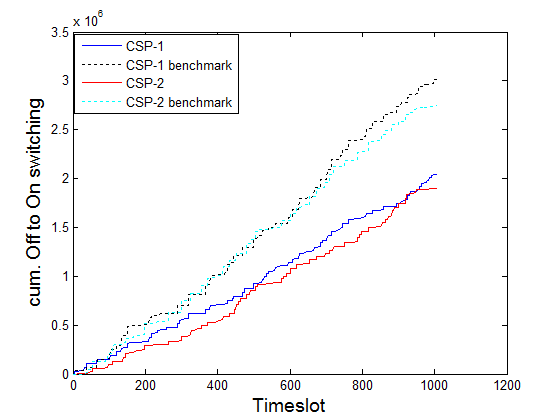
Slower timescale optimization from financial aspect of the CSPs

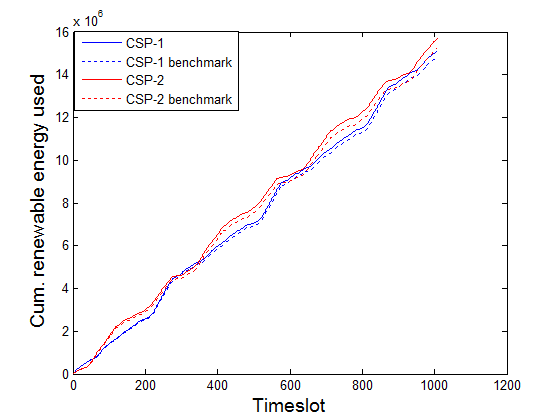


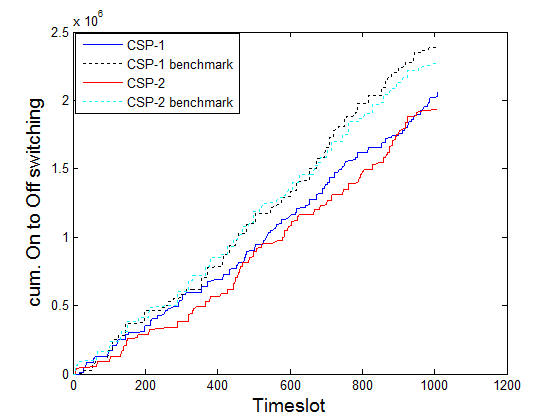


Faster timescale optimization- effects of scheduling









1. Conclusions

In this work we have considered the aggregate cloud service market, consisting of multiple CSPs competing with each other, each aiming to maximizing the revenue.

References

[1] J. Glanz, "Google Details, and Defends, Its Use of Electricity," *The new york times,* Sept, 2011.

[2] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," *ACM SIGCOMM Computer Communication Review,* vol. 39, pp. 123-134, 2009.

[3] A. W. Z. L. I. Liu and H. Mohsenian-Rad, "Opportunities and Challenges for Data Center Demand Response."

[4] M. Ghamkhari, H. Mohsenian-Rad, and A. Wierman, "Optimal Risk-aware Power Procurement for Data Centers in Day-Ahead and Real-Time Electricity Markets."

[5] M. H. Albadi and E. El-Saadany, "A summary of demand response in electricity markets," *Electric Power Systems Research,* vol. 78, pp. 1989-1996, 2008.

[6] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski*, et al.*, "A view of cloud computing," *Communications of the ACM,* vol. 53, pp. 50-58, 2010.

[7] L. A. Barroso and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis lectures on computer architecture,* vol. 4, pp. 1-108, 2009.

[8] B. Jennings and R. Stadler, "Resource management in clouds: Survey and research challenges," *Journal of Network and Systems Management,* pp. 1-53, 2014.

[9] Amazon, "Amazon EC2," [*http://aws.amazon.com/ec2/*](http://aws.amazon.com/ec2/)*,* 2014.

[10] H. Xu and B. Li, "Dynamic cloud pricing for revenue maximization," 2013.

[11] C. Joe-Wong and S. Sen, "Mathematical frameworks for pricing in the cloud: Revenue, fairness, and resource allocations," *arXiv preprint arXiv:1212.0022,* 2012.

[12] R. T. Kaushik, P. Sarkar, and A. Gharaibeh, "Greening the compute cloud's pricing plans," in *Proceedings of the Workshop on Power-Aware Computing and Systems*, 2013, p. 6.

[13] Y. Li, D. Chiu, C. Liu, L. T. Phan, T. Gill, S. Aggarwal*, et al.*, "Towards dynamic pricing-based collaborative optimizations for green data centers," in *Data Engineering Workshops (ICDEW), 2013 IEEE 29th International Conference on*, 2013, pp. 272-278.

[14] R. Pal and P. Hui, "Economic models for cloud service markets," in *Distributed Computing and Networking*, ed: Springer, 2012, pp. 382-396.

[15] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Greening geographical load balancing," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, 2011, pp. 233-244.

[16] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Geographical load balancing with renewables," *ACM SIGMETRICS Performance Evaluation Review,* vol. 39, pp. 62-66, 2011.

[17] A. Krioukov, C. Goebel, S. Alspaugh, Y. Chen, D. E. Culler, and R. H. Katz, "Integrating Renewable Energy Using Data Analytics Systems: Challenges and Opportunities," *IEEE Data Eng. Bull.,* vol. 34, pp. 3-11, 2011.

[18] W. Deng, F. Liu, H. Jin, B. Li, and D. Li, "Harnessing renewable energy in cloud datacenters: opportunities and challenges," *IEEE Network,* vol. 28, pp. 48-55, 2014.

[19] B. Urgaonkar, G. Kesidis, U. V. Shanbhag, and C. Wang, "Pricing of service in clouds: optimal response and strategic interactions," *ACM SIGMETRICS Performance Evaluation Review,* vol. 41, pp. 28-30, 2014.

[20] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems,* vol. 25, pp. 599-616, 2009.

[21] L. Rao, X. Liu, M. D. Ilic, and J. Liu, "Distributed coordination of internet data centers under multiregional electricity markets," *Proceedings of the IEEE,* vol. 100, pp. 269-282, 2012.

[22] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*: Cambridge University Press, 2013.

[23] G. Cook, "Clicking Clean:How companies are creating the green Internet," *Greenpeace,* April, 2014.

[24] S. Boyd and L. Vandenberghe, *Convex optimization*: Cambridge university press, 2004.

[25] A. Krioukov, S. Alspaugh, P. Mohan, S. Dawson-Haggerty, D. E. Culler, and R. H. Katz, "Design and evaluation of an energy agile computing cluster," *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2012-13,* 2012.

[26] A. Subramanian, M. Garcia, A. Dominguez-Garcia, D. Callaway, K. Poolla, and P. Varaiya, "Real-time scheduling of deferrable electric loads," in *American Control Conference (ACC), 2012*, 2012, pp. 3643-3650.

[27] A. Papavasiliou and S. S. Oren, "Supplying renewable energy to deferrable loads: Algorithms and economic analysis," in *Power and Energy Society General Meeting, 2010 IEEE*, 2010, pp. 1-8.

[28] D. E. Bakken, A. Bose, K. M. Chandy, P. P. Khargonekar, A. Kuh, S. H. Low*, et al.*, "GRIP-Grids with intelligent periphery: Control architectures for Grid2050π," in *SmartGridComm*, 2011, pp. 7-12.

[29] E. F. Camacho and C. B. Alba, *Model predictive control*: Springer Science & Business Media, 2013.

[30] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *IEEE computer,* vol. 40, pp. 33-37, 2007.

[31] M. Lin, A. Wierman, L. L. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," *IEEE/ACM Transactions on Networking (TON),* vol. 21, pp. 1378-1391, 2013.

[32] A. Gandhi, V. Gupta, M. Harchol-Balter, and M. A. Kozuch, "Optimality analysis of energy-performance trade-off for server farm management," *Performance Evaluation,* vol. 67, pp. 1155-1171, 2010.

[33] "<http://www.eia.gov/electricity/wholesale/,>" *US energy information administration*.

[34] Z. Liu, I. Liu, S. Low, and A. Wierman, "Pricing data center demand response," in *The 2014 ACM international conference on Measurement and modeling of computer systems*, 2014, pp. 111-123.