To Build or Not to Build? Addressing the Expansion Strategies of Cloud Providers

Data · August 2014

CITATIONS
11

READS
143

4 authors, including:

Mohammad Wardat
Jordan University of Science and Technology
14 PUBLICATIONS **121** CITATIONS
SEE PROFILE

Mahmoud Al-Ayyoub
Ajman University
292 PUBLICATIONS **7,145** CITATIONS
SEE PROFILE

Yaser Jararweh
Jordan University of Science and Technology
383 PUBLICATIONS **8,181** CITATIONS
SEE PROFILE

To Build or Not to Build? Addressing the Expansion Strategies of Cloud Providers

Mohammad Wardat, Mahmoud Al-Ayyoub and Yaser Jararweh
Jordan University of Science and Technology
Irbid, Jordam
Emails: {mawardat12, malayyoub, yaser.amd}@gmail.com

Abdallah A. Khreishah
New Jersey Institute of Technology
NJ, USA
Email: abdallah@njit.edu

***Abstract*—With the increasing popularity gained by cloud computing systems, cloud providers are facing rapidly increasing traffic loads, which requires them to have proper expansion strategies for their datacenters. Expanding the capacities of existing datacenters or building new ones requires many factors to be considered such as power resources availability, prices (of power, land, etc.), carbon tax, free cooling options, and renewable energy. In this paper, we address the problem of deciding the best expansion strategy for a given cloud provider by deciding whether it is beneficial for the cloud provider to build new datacenters or to simply expand the datacenters it currently has.**

***Keywords*—datacenters, renewable power, operation cost, profit maximization, cost minimization, carbon tax, optimization.**

I. Introduction

One of the main concepts related to cloud computing [1], [2] is the move of computations from the user-side to the Internet. With the cloud computing paradigm, companies no longer need to establish and run their own servers to provide on-line services to their customers. Instead, they can simply "rent" the required infrastructure from a specialized cloud provider under a pay-per-use model reducing the Total Cost of Ownership (TCO) and allowing the companies to focus on their own businesses especially in startup companies. Such an option is becoming more appealing for an increasing number of companies, which creates more demand on cloud providers forcing them to optimize their expansion strategies. These expansion strategies should take into consideration both the quality of the service provided to the customers and economic impact on the service provider.

Cloud providers may own several datacenters sparsely distributed across the world to service their clients. Such datacenters are usually huge containing tens of thousands of servers and consuming more power than a medium-size town.[1] Even with these huge datacenters, a cloud provider might still be unable to provide a high quality of service (i.e., one where the service-level agreement (SLA) with the client is not violated) due to the high demand. Thus, expansion strategies must be devised. The cost of expanding a datacenter or building a new one can vary greatly depending on the land cost and the required computing capacity. In this paper, we address the problem of deciding the best expansion strategy for a given cloud provider by deciding whether it is beneficial for the cloud provider to build new datacenters or to simply expand the datacenters it currently has. To solve this problem, one needs to address several issues such as where to build the new datacenters and in which capacities, how to distribute the current and future traffic loads among the new and existing datacenters, etc.

Datacenters are a crucial part for governmental institutions, businesses, industries, and many others. Datacenters vary greatly in size from small in-house datacenters to large scale datacenters that provide their services publicly for millions of users. Datacenters of one service provider may be distributed over a large geographical area which will add an extra overhead in efficiently managing them. In general, datacenters are consuming large amounts of power that my reach up to tens of Megawatts. Also, cooling datacenters are consuming huge amounts of power. These facts are creating many related problems on both the environment and energy resources. A 2010 study showed that large-scale datacenters consumed about 2% of all electricity usage in the United States [3]. This percentage can be translated to be over 100 billion kWh with an approximate cost of $7.4 Billion [4]. Power usage in datacenters is divided into the power consumed by the actual processing system, storage, memory system, etc., and the power consumed by nontechnical components such as ventilation and cooling systems, lighting, etc. The increasing power prices are demanding to reduce the power consumption of datacenter and to increase the usage efficiency of the available power of the datacenters. The new laws for carbon tax are also pushing forward the optimization of power usage. The adoption of renewable energy usage to cover datacenters power requirements is showing a momentum between datacenters owner. Also, building datacenters in locations that provide free air cooling is a good choice for datacenters owner (e.g., Facebook datacenter in Prineville, Oregon). Moreover, management overhead of today datacenters are requiring a lot of man power to handle the extended traffic load. The shortages of such skills is a very serious issue especially in case of constructing many distributed datacenter. Another important issue with having many distributed datacenters is the load balancing between the datacenters. This can be impacted by the availability of high network bandwidth connecting datacenters and its cost.

The rest of this paper is structured as follows. Section II presents the related works. In Section III, we present our system model and evaluate it in Section IV. Finally, we conclude in Section V.

[1] http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html

II. Related Works

The main problem addressed in this work is the expansion strategies of cloud providers to meet the increasing demands of the users. The body of work on this problem is limited since most of the current works focus on optimizing the currently available datacenters by improving power consumption, cooling, request routing, etc. We start our discussion of the related works by discussing these issues before going into the more relevant papers concerned with the added issue of building new datacenters and/or expanding currently available datacenters by increasing its service capacity, which is achieved by increasing the number of servers they contain.

A rich volume of recent research works focus on reducing power costs instead of consumptions. These research works mainly devise different workload distribution policies across geo-distributed datacenters for achieving different performance objectives such as total electricity cost minimization [5], [6], [7], [8], bandwidth cost minimization [9], energy efficiency improvement [10], cooling efficiency [11], reducing carbon footprint [12], maximizing renewable energy usage [13], etc. In order to achieve these objectives researchers have mainly formulated the workload distribution problem as various linear and non-linear optimization problems and adopted various methods and tools to solve it. For optimum solutions, the commonly used mathematical tools are mixed integer programming [6], [7], [8], [14].

Electricity price volatility, [5] is one of the first works to observe the temporal and spatial variability of electricity prices in the wholesale market. The authors argued that as the electricity prices fluctuates across different regions energy expense per unit of computation is not the same for every datacenter. Based on this interesting observation, they designed a distance constrained electricity price optimizer that can achieve significant economic gain. The price optimizer judicially places the load from the client to the datacenter located at cheaper price regions within some radial geographical distance.

The recent research focuses on the future site of the datacenter, because the electricity price is not the same in each region, the datacenter is expected to be built in the location where the lower prices, colder regions and the location where the renewable energy are available ; to reduce the carbon footprint.

In [15], the authors studied the problem of selecting the best locations to build a fixed number of datacenters. They assumed that the datacenters to be built are not associated with any already existing datacenters. They formulated three optimization problems with three objective functions, where the objectives are (i) minimizing the carbon footprint, (ii) minimizing the total cost (including the energy cost, the bandwidth cost and the carbon tax), and (iii) minimizing the average service latency subject to the Quality of Service (QoS) constraints. Their formulation takes as input a set of candidate locations and determines the best location(s) to build future datacenter, the number of servers required at each new datacenter, and how the service requests can be routed to each datacenter.

Another work to select future site(s) is proposed in [16]. The authors introduced a new process that can be used to select the best locations to build new datacenters while considering any already existing datacenters the cloud provider may have. The author formulated an optimization problem considering the following factors: (i) the capital cost (CAPEX), which includes the costs of land acquisition, construction of the infrastructure of datacenter, electricity and bandwidth supplied to the datacenter, etc., (ii) the operational cost (OPEX), which includes the costs of electricity, bandwidth, water for cooling the datacenter, carbon tax, administration staff salary, etc., (iii) response time, which depends on the distance between the location of datacenter and a population center, (iv) consistency delay, which depends on the distance between two potential locations of the datacenter and (v) availability. The main objective of the formulated problem is to minimize the total cost (i.e., CAPEX & OPEX) subject to response time, consistency delay and availability.

Reducing the carbon footprint and maximizing renewable energy usage are the objectives considered by the authors of [17] who showed that the carbon footprint can be reduced by building the datacenters near to the sites of the renewable energy. They developed a mathematical model for calculating the total carbon footprint including: (i) the manufacturing footprint (ii) the usage footprint and (iii) the communication footprint. The authors discussed how to reduce the carbon footprint (especially, the manufacturing carbon footprint) by redistributing the load of datacenter to other sites as the availability of renewable energy.

The authors of [6] formulated a model to address the total electricity cost problem under diverse electricity prices across different regions and time periods while maintaining QoS guarantees. This model for total electricity cost, workload load, end-to-end delay constraint for datacenters. They formulated the model as a mixed-integer programming problem where the constraints captured the workload requirements and the service delay assurances. The authors solved it using the fast polynomial algorithm proposed by Brenner [18].

Compared with the previously mentioned related works, the contribution of this work lies in the following points. No prior work has addressed the problem of determining the best future location of the datacenter while taking into account the tradeoff between maximizing the revenue and minimizing the operational cost of the datacenter instead of (bandwidth, cooling, carbon tax, and power costs). Moreover, previous works neglect important economical aspects such as the annual inflation in the costs (bandwidth, cooling, price of electricity) and in the revenue.

III. System Model

In this section, an optimization problem is formulated using mixed integer programming to address the problem of determining the best expansion strategy a cloud provider can take to face the increasing demands and increase its revenue. The computed strategy may include expanding current datacenters by increasing the number of servers they contain or building new datacenters (which involves determining how many datacenters to build, where to build them and in which capacities). This is achieved by calculating the profit gained in each year of the period under consideration. Taking a look at the accumulated and inflated profits over the years and comparing it with what would the initial investment

gain by placing it in a bank makes the decision of whether to build new and/or expand the current datacenters an easy decision. Inflation is discussed in subsection III-A, whereas in subsection III-B, we present an extension of the proposed model that takes into account the effect of renewable energy more explicitly.

In the proposed model, the cloud provider inputs its current datacenter locations along with the number of servers each one has and the system will find the revenue maximizing option regarding building new and/or expanding the current datacenters. We consider a discrete-time model, in which the time period of interest is discretized on two levels: a major level and a minor level. On the major level, the overall time period is divided into T time segments, where each segment can represent a decade, a year, a month, etc., while, on the minor level, each major time segment is divided into H timeslots. In the following year, we consider T and H to be the number of years and the number of hours in each year, respectively.

Before describing our model, we briefly go over the notations used and the assumptions made. In order for our model to work, we have to specify discrete sets of user locations,[2] denoted by U, and datacenter locations, denoted by S, which includes the set of current locations along with the set of candidate location on which the cloud provider can build datacenters. Now, we define a set of binary variables, $X = \{x_s^t | t \in T, s \in S\}$,[3] to denote whether a datacenter is built on location s at year t. Obviously, we must make sure that if a datacenter is built on certain location in a certain year, it stays like this for the following years (i.e., if $x_s^{t_1} = 1$ then $x_s^{t_2}$ must also be 1 for all $t_2 > t_1$). Currently built datacenters are easy to handle in this way. If s is the location of a currently built datacenter, then $x_s^t = 1$ for all $t \in T$.

By taking into account the change in the price of electricity in different locations at different times during the day, the authors of [15] proposed a request distribution policy to route the parts service requests to potentially different datacenters. For this purpose, we denote the total number of service requests originating from user location u during hour h of year t by $L_u^{t,h}$ and the portion of $L_u^{t,h}$ serviced by the datacenter s in location s by $\lambda_{s,u}^{t,h}$. Let $\Lambda = \{\lambda_{s,u}^{t,h} | s \in S, u \in U, t \in T, h \in H\}$. The following constraint ensures that no request is denied.

$$\sum_{s \in S} \lambda_{s,u}^{t,h} = L_u^{t,h}, \quad \forall h \in H, t \in T \tag{1}$$

We mow define a binary variable ($y_{s,u}^{t,h}$) to represent the ability of datacenter s to handle service requests from user location u at hour h in year t. Let $Y = \{y_{s,u}^{t,h} | s \in S, u \in U, t \in T, h \in H\}$. Obviously, if a datacenter is not yet built at a certain location, it cannot service any request. Thus, we have the following constraint.

$$y_{s,u}^{t,h} \leq x_s^t, \quad \forall s \in S, u \in U, h \in H, t \in T \tag{2}$$

[2]User locations could be cities, towns, etc.

[3]For the sake of simplifying the presentation, we are slightly abusing the notation and treat T, H, U, S and X as both sets (representing the years of the project, the hours in each year, the user locations, the datacenter locations and the binary variables representing whether a datacenter is built on a certain location in a certain year, respectively) and integers (representing the sizes of these sets).

Moreover, to ensure that if a datacenter s does not receive a service request it is not ready to handle, we use the following constraint.

$$0 \leq \lambda_{s,u}^{t,h} \leq y_{s,u}^{t,h} L_u^{t,h}, \quad \forall s \in S, u \in U, h \in H, t \in T \tag{3}$$

We define m_s^t to be the number of servers in datacenter s during year t. Let $M = \{m_s^t | t \in T, s \in S\}$. The number of servers in any datacenter is bounded by lower and upper bounds represented by $M^{\min}$ and $M^{\max}$, respectively. Then we have:

$$x_s^t M^{\min} \leq m_s^t \leq x_s^t M^{\max}, \quad \forall s \in S \tag{4}$$

The total power consumption in the datacenter id divide into two types depending on whether power is consumed by an IT equipment (such as servers, routes, etc.) or not (e.g., for conversion, lighting, and cooling, etc.). The ratio between the total power consumption to the IT equipment power consumption is donated by E_{usage} and it is used as a measure for a datacenter's power usage efficiency (PUE) [19]. As for the power consumption of the servers, we denote the average power consumption of a single server when the server is idle by P_{idle} and when it is handling the service request by P_{peak}. Following the model of [20], we can calculate the power consumption in candidate location s for a certain hour h in year t as follows.

$$\begin{aligned} P_s^{t,h} &= m_s^t (P_{\text{idle}} + (E_{\text{usage}} - 1) P_{\text{peak}}) \\ &+ m_s^t (P_{\text{peak}} - P_{\text{idle}}) \gamma_s^{t,h} + x_s^t \epsilon \end{aligned} \tag{5}$$

where ϵ is an empirically derived constant and $\gamma_s^{t,h}$ denotes the average server utilization of the datacenter s during hour h of the year t defined as:

$$\gamma_s^{t,h} = \frac{\sum_{u \in U} \lambda_{s,u}^{t,h}}{m_s^t \mu} \tag{6}$$

where μ denotes the total number of service requests that a computer server can handle in one hour. Note that although the last two equations seem non-linear, they can be easily linearized by plugging the definition of [19] into Equation 5.

Datacenters require so much energy. In fact, some large datacenters consume more power than a medium-size town. The power plant in each region supplies the power for subscribers, commercial, residential, and industrial load, which leads to varying demand throughout the day. Moreover, some of the power plants depend on renewable energy sources such as the wind, the sun, etc. So, the proposed model takes into account the amount of available power at each hour of the day as follows.

$$P_s^{t,h} \leq P_s^{t,h,\max}, \quad \forall s \in S, h \in H, t \in T \tag{7}$$

Several factors affect the quality of the provided service and may cause violations in the service-level agreement (SLA). Delay is one of these factors. Different types of delay have been explored in the literature. In this model, we focus only on the propagation delay. The following constraint makes sure that the the propagation delay for any request from user u serviced by datacenter s (denoted $D_{s,u}$) does not exceed the maximum delay allowed by the SLA.

$$2 D_{s,u} y_{s,u}^{t,h} \leq D^{\max}, \quad \forall s \in S, u \in U, h \in H \tag{8}$$

In order to avoid other SLA violations, we limit the average server utilization at each datacenter by a constant upper bound $\gamma^{\max} \in (0, 1]$. Thus, we have the following condition.

$$\gamma_s^{t,h} \leq \gamma^{\max}, \quad \forall s \in S, h \in H \tag{9}$$

The value of $\gamma^{\max}$ depends on the quality of service and the type of service request. In this model, the type of service we consider is web service request, therefore, the value $\gamma^{\max}$ is small enough to avoid waiting time.

Now, we are ready to present our formulation. The input parameters include the set of user locations, the set of the datacenter locations, the hourly traffic loads from each user location, the propagation delay between each user location and each datacenter location along with the upper bound on the propagation delay, the power consumptions of a single server when it is idle and when it is processing a request, and, for each datacenter location, the PUE, the hourly power constraint, the maximum utilization and the maximum and minimum capacities (in terms of the number of servers) of each datacenter. The parameters to be computed are X, M, Y and Λ. Note that the sets X and M might be partly filled with information about currently built datacenters as follows. If s is the location of a currently built datacenter, then $x_s^t = 1$ for all $t \in T$, and m_s^0 is set to the number of servers already in datacenter s. The formulation is as follows.

$$\begin{aligned} \text{Maximize}_{x,m} \quad & \text{RV}(T) - (\text{OPEX}(T) + \text{CAPEX}(T)) \\ \text{Subject to} \quad & \text{Constraints} \quad 1 - 9. \end{aligned}$$

The notation will be explained in the following paragraphs.

The overall cost of the datacenters can be divided into operational cost (OPEX) and capital cost (CAPEX). CAPEX includes the costs of land acquisition, construction of the infrastructure of datacenter, electricity and bandwidth supplied to the datacenter, etc., whereas OPEX includes the costs of electricity, carbon tax, bandwidth cost, etc. More formally, CAPEX for a certain year t can be expressed using the following equation.

$$\text{CAPEX}(t) = \sum_{s \in S} (x_s^{t-1} - x_s^t)\text{BC}_s^t + (m_s^{t-1} - m_s^t)\text{SC}_s^t,$$

where BC_s^t represent the cost of building a datacenter s in year t and $\text{SC}_{t,s}$ represent the cost of buying a server for the datacenter s in year t.

To maximize the profit, cloud providers are interested in reducing OPEX, which means that locations with low electricity prices are favorable. However, choosing such locations might not be the most environmentally responsible decision. For example, in Wyoming and Utah the price of electricity is cheaper, because of their coal-fired power plants [15]. The carbon footprint of coal-fired and natural gas generators is higher than nuclear and hydroelectric generators [21]. OPEX for a certain year t can be expressed as follows [15].

$$\text{OPEX(t)} = \sum_{s \in S} \sum_{h \in H} (\theta_s^t P_s^{t,h} + \delta_s^t(\rho_s + 1) P_s^{t,h} + \sum_{u \in U} (\lambda_{s,u}^{t,h} \sigma_{s,u}^t)),$$

where δ_s^t is the carbon tax in location s in year t, ρ_s is the power transmission loss rate location s, $\sigma_{s,u}^t$ is the cost of the bandwidth between user location u and candidate location s and θ_s^t is the price of electricity in candidate location s taken during three different time-of-use price periods: off-peak (when the demand for electricity is low), mid-peak (when the demand for electricity is moderate; generally, during daytime, but not the busiest times of day) and on-peak (when the demand for electricity is high; generally, when people are cooking, firing up their computers and running heaters or air conditioners).

Now, the revenue of year t is computed using the following equation [22]: $\text{RV}(t) = ((1 - p(x))\alpha^t \lambda_{s,u}^{t,h} - p(x)\beta^t)$, where $p(x)$ is the probability that the waiting time for a service request exceeds the SLA-deadline, α^t is the service fee that the datacenter charges the costumers for handling a single service request and β^t is the penalty that the datacenter must pay for every service request it cannot handle (thus, causing an SLA violation).

A. Inflation

Due to insufficient data, several input parameters (such as the traffic loads) cannot be predicted accurately. The best we can do is to compute the current (or past) values for such parameters and "inflate" them as shown in the following paragraphs. Inflation is also important since the time interval considered in this model may span several years and we need to predict future monetary values of certain things (such as electricity). Moreover, any amount of money (whether it is a profit or a loss) setting for any amount of time (months, years, etc.) must be inflated. In this work, several values are inflated such as the traffic loads $(L_u^{h,t})$, the electricity prices (θ_s^t), the carbon taxes (δ_s^t), the bandwidth costs $(sigma_{s,u}^t)$, service fees (α^t), penalties for SLA violations (β^t). the initial investment and the yearly revenue. Of course, these different values might require different inflation rates. In our simulation results, we try to use realistic values for these rates based on our reading of the literature.

To handle these cases, we define the following functions. We start with the compound interest, which can be computed as $A = P\left(1 + \frac{r}{n}\right)^{nt}$, where A is the amount of money accumulated after t years, including interest, P is the principal amount (the initial amount you borrow or deposit), r is the annual rate of interest (as a decimal), t is number of years the amount is deposited or borrowed for and n is the number of times the interest is compounded per year. Now, we move to the Compound Annual Growth Rate (CAGR), which is the interest rate at which a given present value would "grow" to a given future value in a given amount of time. The formula for computing CAGR is: $\left(\frac{\text{FV}}{\text{PV}}\right)^{\frac{1}{t}} - 1$, where FV is the future value, PV is the present value and t is the number of years. Finally, the formula for the inflation rate is: $P_n = P(1 + i)^n$, where P_n is total inflated/estimated cost, i is the inflation rate and t is the difference between the base year and the selected year. Alternatively, we can use the following simpler (linear) to compute inflation $P_n = P + (P \times i \times n)$.

B. Renewable Energy

The model discussed so far does not explicitly account for renewable energy, which is one of the biggest concerns related to datacenters and their effect on the surrounding environment. To address this issue, we reformulate Equation 5

TABLE I. THE NUMBER OF SERVERS IN EACH DATACENTERS DURING 5 YEARS.

Year	1	2	3	4	5
DC1	5000	5000	15122	31411	46784
DC2	0	5000	30237	50000	50000
DC3	0	0	0	0	5000
DC4	41677	47097	49999	50000	50000
DC5	41677	50000	50000	50000	50000
DC6	0	0	0	0	6971
DC7	0	5000	15122	31402	50000
DC8	0	5000	15123	31394	50000
DC9	41677	50000	50000	50000	50000
DC10	0	0	0	12797	50000
DC11	36678	47097	50000	50000	50000
DC12	41677	50000	50000	50000	50000
Total	208386	264194	325603	407004	508755

as follows [22].

$$\begin{aligned} P_s^{t,h} &= [m_s^t(P_{\text{idle}} + (E_{\text{usage}} - 1)P_{\text{peak}}) \\ &+ m_s^t(P_{\text{peak}} - P_{\text{idle}})\gamma_s^{t,h} + x_s^t\epsilon - x_s^t G_s^{t,h}]^+, \end{aligned}$$

where $[x]^+ = \max\{x, 0\}$ and $G_s^{t,h}$ is the amount of renewable power generated in location s during hour h of year t. The amount of power exchange with the power grid is obtained as $[P-G]$. If local renewable power generation is lower than local power consumption, i.e., $P > G$, then $[P-G]$ is positive and the power flow is in the direction from the power grid to the datacenter. If $P = G$ then the datacenter operates as a zero-net energy facility. Now, if $P < G$, then $[P-G]$ is negative and the power flow is in the direction from the datacenter to the power grid [22]. Note that, $[P-G]^+$ indicates the amount of power to be purchased from the grid. If this term is negative, the datacenters electricity cost will be zero, given the assumption that the grid does not provide compensation for the injected power [22]. In the simulations, we were forced to ignore this important extension due to the lack of realistic input data for the different types of renewable energy generators.

IV. EXPERIMENTS AND RESULTS

In this section, we discuss the simulation experiments we conducted on our system and the obtained results. We start by discussing the candidate locations. We focus on contiguous US for simplicity and due to the fact that most of the required data is available for this part of the world. Since the power availability is limited in certain regions, we need to exclude states generating power at rates smaller than their consumptions. According to [23], the excluded states are CA, NV, ID, SD, MN, WI, OH, TN, FL, NC, VA, MD, NY, DE, NJ, CT, RI, VT, MA, and DC. As for the remaining states, we must make sure that Constraint 7 is satisfied, so, we compute the maximum available power in each state as follows. For IA, KY and MS, the maximum available power is 60 megawatts while other states such as WA, NH, OR, OK, UT, WY, IL, AZ, PA and SC can handle larger demands (greater then 100 megawatts).

After deciding the set S, we turn our attention to other input parameters. According to [20], we set P_{peak} and P_{idle} to 140 and 84 watts. A fixed value of 2 is a common choice in the literature for PUE [19]. However, we do consider a more realistic case where the PUE changes with varying outside temperature as shown in Figure 2. For the sake of simplicity, we consider only four different outside temperatures for each location depending on whether the considered time is in the Summer or the Winter season and whether it is during daytime or at night.

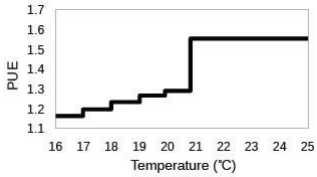


Fig. 2. PUE values for different outside temperature [16].

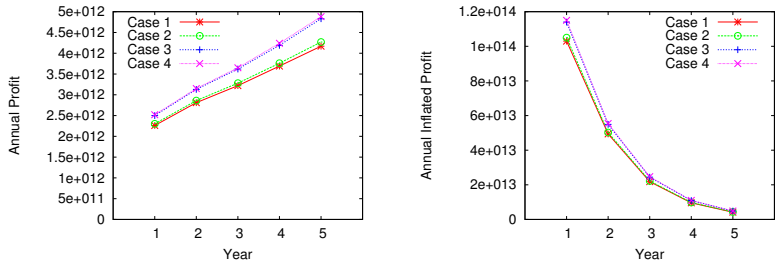
TABLE II. TIME PERIODS FOR BOTH VARYING PUE VALUES AND DYNAMIC PRICING.

period	Summer May 1 - October 31	Winter November 1 - April 30
Daytime	6am-7pm	7am-6pm
Night	7pm-6am	6pm-7am
Off-peak	7pm-7am	7pm-7am
Mid-peak	7-11am & 5-7pm	11am-5pm
On-Peak	11am-5pm	7-11am and 5-7pm

The temperatures are taken from online weather websites such as weatherbase.com and worldweatheronline.com. The details of these periods are shown in Table II.

As for the traffic load, we choose the total number of service requests incoming from all user locations to be between 1.5 and 2 million hits/sec [5]. We assume that each server can process one request per second, i.e., $\mu = 3600$. We set $\gamma^{\max} = 0.8$ [24]. The electricity price information based on the average price for industrial load is available [25]. As mentioned in Section III, we consider three different time-of-use price periods: on-peak, off-peak and mid-peak. Moreover, we assume two different seasons: Winter and Summer. The price of electricity vary from one period to the other by as much as 3 cent/kWh. Table II shows the details of the considered time periods. Finally, we note that we focus in our simulation on OPEX and exclude CAPEX due to the limited publicly available information about the input parameters required by CAPEX.

Now, let us present and discuss the results of the two experiments we conduct. The objective of the first experiment is to study the decisions made by the proposed formulation regarding the best expansion strategies to handle the increasing traffic load. We run our model for five years on 12 datacenter locations, half of which have already built datacenters. Remember that an already built datacenter will have some servers in it. Since the servers are homogenous and resources (like servers) are only added when needed, the number of servers in each datacenter is an indication of how much traffic load it is processing. We assume that the maximum number of servers to be placed in a single datacenter is 50,000. Table I shows how the number of servers in each datacenter increases with the passage of time and the increase in the traffic load. At the beginning, only one (DC1) out of the six already built datacenters was lightly loaded small while the other five were heavily loaded. In the second year, the lightly loaded datacenter (DC1) remained lightly loaded (probably due to its high operational cost or high delay compared with the other available datacenters) and the heavily loaded datacenters almost reached their full capacity. Moreover, three new datacenters were built. The same trend continues in the following years. Datacenters with low operational cost or low delay expand in terms of the number of servers until they reach their full capacities. If this is not enough to process the newly generated traffic, either new datacenters are built or the datacenters with high operational cost are expanded depending on which option provides better profits. By the last year of this experiment, the cloud provider



(a) Original (non-inflated) profits.

(b) Inflated profits.

Fig. 1. The annual profits for the four cases under consideration.

is forced to build datacenters in all location to process the huge amount of traffic load.

In the second experiment, we study the effect of using fixed PUE value vs varying PUE values as well as using flat rate electricity pricing vs dynamic pricing. Thus, the four cases under consideration are: (1) fixed PUE and flat rate prices, (2) fixed PUE and dynamic prices, (3) varying PUE and flat rate prices and (4) varying PUE and dynamic prices.

Figures 1(a) and 1(b) show the annual profits (original and inflated) generated for the four cases under consideration.The effect of inflation (an issue usually ignored in many related works) is obvious in the two figures. While Figure 1(a) shows a significant increase in the actual gained profits of each year, Figure 1(b) shows an opposite trend for the inflated profits since the profits made in the first year is exposed to inflation for a longer period of time which makes them much higher than profits made in the last year, which were not inflated at all.

From Figure 1(a), it can be seen that using different PUE values for different times of the day generates (an average of 2%) better annual profits than using fixed PUE value. Moreover, using dynamic pricing also has even more positive effect on the annual profits as it increases them by an average of 13%. Finally, Mixing both dynamic settings (varying PUE values and dynamic pricing) causes an average improvement of 14% on the annual profits. Similar trends are shown in Figure 1(b) for the annual inflated profits.

V. Conclusion

In this work, we addressed the problem of deciding the best expansion strategy for a given cloud provider by deciding whether it is beneficial for the cloud provider to build new datacenters or to simply expand the datacenters it currently has. We proposed a formulation of the problem that takes into account the locations and capacities of the future datacenters, the operational cost of the datacenters, important economical aspects such as the annual inflation in the costs and revenue. Traffic and computing resources heterogeneity could increase the room for better optimizing the usage of today datacenter. We will be considering this as a future work to extend this paper.

References

[1] J. Shi, M. Taifi, and A. Khreishah, "Resource planning for parallel processing in the cloud," in *IEEE HPCC*, 2011.

[2] J. Shi, M. Taifi, A. Khreishah, and J. Wu, "Sustainable gpu computing at scale," in *IEEE CSE*, 2011.

[3] J. Koomey, "Growth in data center electricity use 2005 to 2010," Analytics Press, Oakland, CA, 2011.

[4] A. Rahman, X. Liu, and F. Kong, "A survey on geographic load balancing based data center power management in the smart grid environment," *IEEE Communications Surveys & Tutorials*, 2014.

[5] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," *SIGCOMM Computer Communication Review*, 2009.

[6] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing electricity cost: optimization of distributed internet data centers in a multi-electricity-market environment," in *IEEE INFOCOM*, 2010.

[7] ——, "Coordinated energy cost management of distributed internet data centers in smart grid," *IEEE TSG*, 2012.

[8] Y. Zhang, Y. Wang, and X. Wang, "Capping the electricity cost of cloud-scale data centers with impacts on power markets," in *HPDC*, 2011.

[9] N. Buchbinder, N. Jain, and I. Menache, "Online job-migration for reducing the electricity bill in the cloud," in *NETWORKING*, 2011.

[10] J. Li, Z. Li, K. Ren, and X. Liu, "Towards optimal electric demand management for internet data centers," *IEEE TSG*, 2012.

[11] K. Le, R. Bianchini, J. Zhang, Y. Jaluria, J. Meng, and T. D. Nguyen, "Reducing electricity cost through virtual machine placement in high performance computing clouds," in *SC*, 2011.

[12] J. Doyle, D. O'Mahony, and R. Shorten, "Server selection for carbon emission control," in *SIGCOMM Workshop on Green Networking*, 2011.

[13] R. Carroll, S. Balasubramaniam, D. Botvich, and W. Donnelly, "Dynamic optimization solution for green service migration in data centres," in *IEEE ICC*, 2011.

[14] Z. Abbasi, T. Mukherjee, G. Varsamopoulos, and S. K. Gupta, "Dynamic hosting management of web based applications over clouds," in *HiPC*, 2011.

[15] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Energy-information transmission tradeoff in green cloud computing," *Carbon*, 2010.

[16] Í. Goiri, K. Le, J. Guitart, J. Torres, and R. Bianchini, "Intelligent placement of datacenters for internet services," in *ICDCS*, 2011.

[17] W. Van Heddeghem, W. Vereecken, D. Colle, M. Pickavet, and P. Demeester, "Distributed computing for carbon footprint reduction by exploiting low-footprint energy availability," *FGCS*, 2012.

[18] U. Brenner, "A faster polynomial algorithm for the unbalanced hitchcock transportation problem," *Operations Research Letters*, 2008.

[19] R. Brown *et al.*, "Report to congress on server and data center energy efficiency: Public law 109-431," 2008.

[20] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *SIGARCH Computer Architecture News*, 2007.

[21] S. Baldwin, "Carbon footprint of electricity generation," London: Parliamentary Office of Science and Technology, 2006.

[22] M. Ghamkhari and H. Mohsenian-Rad, "Energy and performance management of green data centers: A profit maximization approach," *IEEE TSG*, 2013.

[23] Energy Information Administration, "State electricity profiles - summer capacity," 2008.

[24] K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiran, F. Tobagi, and C. Diot, "Analysis of measured single-hop delay from an operational backbone network," in *IEEE INFOCOM*, 2002.

[25] Energy Information Administration, "Average retail price of electricity to ultimate customers by end-use sector by state," 2010.