Optimizing Expansion Strategies for Ultrascale Cloud Computing Data Centers

Article *in* Simulation Modelling Practice and Theory · March 2015

CITATIONS
7

READS
197

4 authors, including:

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

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

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

ARTICLE IN PRESS

Simulation Modelling Practice and Theory xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

Optimizing expansion strategies for ultrascale cloud computing data centers

Mahmoud Al-Ayyoub [a,*], Mohammad Wardat [a], Yaser Jararweh [a], Abdallah A. Khreishah [b]

[a] Department of Computer Science, Jordan University of Science and Technology, Irbid 22110, Jordan
[b] New Jersey Institute of Technology, NJ, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Ultrascale data centers
Cloud computing
Expansion modeling
Traffic loads modeling
Optimization problems
Mixed integer-linear programming

ABSTRACT

With the increasing popularity gained by cloud computing systems over the past few years, cloud providers have built several ultrascale data centers at a variety of geographical locations, each including hundreds of thousands of computing servers. Since cloud providers are facing rapidly increasing traffic loads, they must have proper expansion strategies for their ultrascale data centers. The decision of expanding the capacities of existing data centers or building new ones over a certain period requires considering many factors, such as high power consumption, availability of resources, prices (of power, land, etc.), carbon tax, free cooling options, and availability of local renewable power generation. While a rich volume of recent research works focused on reducing the operational cost (OPEX) of the data centers, there exists no prior work, to the best of our knowledge, on investigating the trade-off between minimizing the OPEX of the data centers and maximizing their revenue from the services they offer while respecting the service level agreement (SLA) with their customers. In this study, we model this optimization problem using mixed integer-linear programming. Our proposed model is unique compared to the published works in many aspects such as its ability to handle realistic scenarios in which both data centers' resources (servers) and user generated traffic are heterogeneous. To evaluate the proposed model and the impact of different parameters on it's performance, several simulation experiments are conducted.

1. Introduction

One of the main concepts related to cloud computing is the migration 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 the case of 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 the economical impact on the service provider [1,2].

* Corresponding author.
E-mail addresses: maalshbool@just.edu.jo (M. Al-Ayyoub), mawardat12@gmail.com (M. Wardat), yijararweh@just.edu.jo (Y. Jararweh), abdallah@njit.edu (A.A. Khreishah).

http://dx.doi.org/10.1016/j.simpat.2015.03.002
1569-190X/

Cloud providers may own several data centers distributed across different locations to serve their clients. Such data centers are usually huge containing tens of thousands of servers and consuming more power than a medium-size town.[1] Even with these huge data centers, 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 data center 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 data centers or to simply expand the data centers it currently has. To solve this problem, one needs to address several issues such as where to build the new data centers and in which capacities and how to distribute the current and future traffic loads among the new and existing data centers.

Data centers are a crucial part for governmental institutions, businesses, industries, and many others. They vary greatly in size from small in-house data centers to large scale data centers that provide their services publicly for millions of users. Data centers of one service provider may be distributed over a large geographical area which requires an extra overhead for managing them efficiently. Moreover, they consume large amounts of power that can reach up to tens of megawatts for running their hardware and cooling them. These facts are creating many problems on both the environment and energy resources. A 2010 study showed that large-scale data centers consumed about 2% of all electricity usage in the United States [3]. This percentage can be translated to be over 100 billion kW h with an approximate cost of $7.4 billions [4]. Power usage in data centers is divided into the power consumed by the IT components and the power consumed by non-IT components such as ventilation and cooling systems, and lighting.

Being environmentally responsible is definitely a concern in the cloud computing society. Researchers from both Academia and the industry are collaborating to address environment grand challenges and to accelerate the research in this field [5]. Managing carbon footprint and power consumption [6] are examples of such efforts. From a monetary perspective, the increasing prices of power offer more reasons to reduce the power consumption of data center and to increase the usage efficiency of the available power. The new laws for carbon tax are also pushing forward the optimization of power usage. The adoption of renewable energy usage to cover data centers power requirements is showing a momentum between data centers owners. Also, building data centers in locations that provide free air cooling is a good choice for data centers owners (e.g., Facebook data center in Prineville, Oregon). Moreover, management overhead of today's data centers requires a lot of manpower to handle the extended traffic loads. The shortage of such skills is a very serious issue especially in case of constructing many distributed data centers. Another important issue with having many distributed data centers is the load balancing between the data centers. This can be impacted by the availability and cost of high network bandwidth connecting data centers.

The contributions of our work are as follows. First, the objective of our proposed model is to decide the best expansion strategy for a given cloud provider by deciding whether it is beneficial for the cloud provider to build new data centers or to simply expand the data centers it currently has. To the best of our knowledge, no prior work has addressed this problem explicitly. Second, our proposed model addresses the problem of heterogeneity of resources (like servers) and traffic types (with their varying delay constraints). This is another aspect that has not been addressed explicitly before, to the best of our knowledge. Third, our proposed model aims to satisfy the conflicting goals of maximizing the revenue while minimizing the operational cost (OPEX) for the provider. Moreover, it has to perform well for varying conditions at the different geographical regions, varying prices of electricity, different kinds of renewable power sources and their availabilities and different traffic types throughout the day/year.

The rest of this paper is organized as follows. Section 2 discusses the system model. Section 3 explains the simulation results and shows the optimization results. Section 4 includes a literature review for some of the optimization techniques. Finally, the conclusion and future work are presented in Section 5.

2. System model

In this section, an optimization problem is formulated using mixed integer-linear programming to address the problem of determining the best expansion strategy a cloud provider can take to face the increasing demands and to increase its revenue. The computed strategy may include expanding current data centers by increasing the number of servers they contain or building new data centers (which involves determining how many data centers to build, where to build them and in which capacities). As part of the solution, the formulation also addresses the problem of how to distribute the service request among the data centers to achieve the highest revenue. The proposed model achieves its goal by calculating the profit gained in each year of the period under consideration. Taking a look at the accumulated and inflated profit over the years and comparing it with what would the initial investment gain (e.g., by placing it in a savings account or in bonds) makes the decision of whether to build new and/or expand the current data centers an easy decision. Fig. 1 represents our system model.

This section covers many issues. The expansion strategy optimization model is discussed in Section 2.1, whereas in Section 2.2, we present the heterogeneity of resources and traffic model. Inflation is discussed in Section 2.3. In Section 2.4, we present an extension of the proposed model to take into account the effect of renewable energy more explicitly.

[1] http://goo.gl/zg2PWg.

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

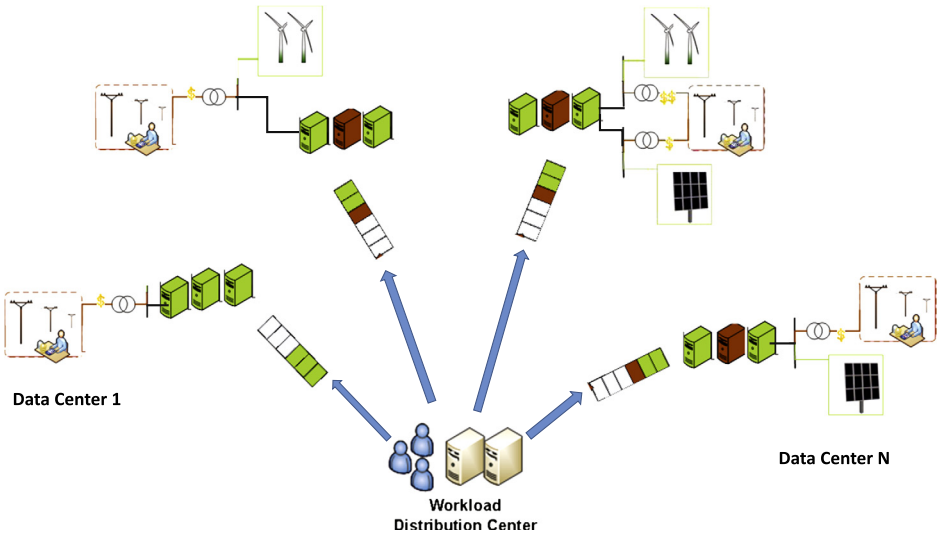

Fig. 1. System model for ultrascale cloud computing data centers.

2.1. The expansion strategy optimization model

An earlier version of this proposed model appeared in [7] where the cloud provider inputs its current data center locations along with the number of servers each one has and then the system will find the revenue maximizing option regarding building new and/or expanding the current data centers. We consider a discrete-time model, in which the time period of interest is discretized in 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, 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, denoted by U, which could be cities, towns, etc., and data center locations, denoted by S, which includes the set of existing data center locations along with the set of candidate locations on which the cloud provider can build new data centers. Now, we define a set of binary variables, $X = \{x_s^t | t \in T, s \in S\}$, to denote whether a data center is built on location s at year t. Obviously, we must make sure that if a data center is built on a 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 $t_2 > t_1$). Existing data centers are easy to handle in this way. If s is the location of an existing data center, then $x_s^t = 1$ for all $t \in T$.

By taking into account the changes in electricity price in different locations at different times during the day, the authors of [8] proposed a request distribution policy to route the parts of service requests to potentially different data centers. 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 $L_u^{t,h}$ served by the data center at 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 define a binary variable, $y_{s,u}^{t,h}$, to represent the ability of data center at location 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 data center is not yet built at a certain location, it cannot service any request. Thus, we have the following constraint.

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

Moreover, to ensure that a data center at location s does not receive a service request it is not ready to handle, we use the following constraint.

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

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

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

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

The total power consumption in the data center is divided into two types depending on whether power is consumed by an IT equipment (such as servers and routers) 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 data center's power usage efficiency (PUE) [9]. 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 [10], we can calculate the power consumption in candidate location s for a certain hour h in year t as follows:

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

where ϵ is an empirically derived constant and $\gamma_s^{t,h}$ denotes the average server utilization of the data center at location 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 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 Eq. (6) into Eq. (5), then we reformulate Eq. (5) as following:

$$P_s^{t,h} = m_s^t(P_{\text{idle}} + (E_{\text{usage}} - 1)P_{\text{peak}}) + (P_{\text{peak}} - P_{\text{idle}})\left(\frac{\sum_{u\in U}\lambda_{s,u}^{t,h}}{\mu}\right) + x_s^t\epsilon$$

The power plant in each region supplies power to different types of subscribers (commercial, residential, and industrial) which leads to varying demand throughout the day. Moreover, some of the power plants depend on renewable energy sources such as the wind and the sun. So, the proposed model must enure that the total amount of consumed power does not exceed the amount of available power at any hour.

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

Several factors affect the quality of the provided service and may cause violations in the 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 propagation delay for any request from user u served by data center at location s, denoted by $D_{s,u}$, does not exceed the maximum delay allowed by the SLA.

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

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

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

The value of $\gamma^{\max}$ depends on the quality of service and the type of service request. For now, the type of service we consider is web service request, therefore, the value $\gamma^{\max}$ is small enough to avoid waiting time. Section 2.2 presents an extension of this model capable of handling heterogenous service requests.

Now, we are ready to present our formulation. The input parameters include the set of user locations, the set of the data center locations, the hourly traffic loads from each user location, the propagation delay between each user location and each data center location along with the upper bound on the propagation delay, the power consumption of a single server when it is idle and when it is processing a request, and, for each data center 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 data center. 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 existing data centers as follows. If s is the location of an existing data center, then $x_s^t = 1$ for all $t \in T$, and m_s^0 is set to the number of servers already in data center at location s. The formulation is as follows.

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

The notation is explained in the following paragraphs.

The overall cost of the data centers can be divided into operational cost (OPEX) and capital cost (CAPEX). CAPEX includes the costs of land acquisition, construction of the infrastructure of the data center, electricity and bandwidth supplied to the data center, 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.

$$\text{CAPEX}(t) = \sum_{s\in S}(x_s^t - x_s^{t-1})\text{BC}_s^t + (m_s^t - m_s^{t-1})\text{SC}_s^t, \tag{10}$$

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

where BC_s^t represent the cost of building a data center at location s in year t and SC_s^t represent the cost of buying a server for the data center at location s in year t. Berral et al. [11] present a more detailed equation for CAPEX.

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 cheap because of their coal-fired power plants [8]. The carbon footprint of coal-fired and natural gas generators is higher than nuclear and hydroelectric generators [12]. OPEX for a certain year t can be expressed as follows [8].

$$\mathrm{OPEX(t)} = \sum_{s\in S}\sum_{h\in H}\left(\theta_s^t P_s^{t,h} + \delta_s^t(\rho_s+1)P_s^{t,h} + \sum_{u\in U}(\lambda_{s,u}^{t,h}\sigma_{s,u}^t)\right) \tag{11}$$

where δ_s^t is the carbon tax in location s in year t, ρ_s is the power transmission loss rate at 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 [13].

$$\mathrm{RV}(t) = ((1-p(x))\alpha^t\lambda_{s,u}^{t,h} - p(x)\beta^t)$$

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 data center charges the customers for handling a single service request and β^t is the penalty that the data center must pay for every service request it cannot handle (thus, causing an SLA violation).

2.2. The heterogeneity of resources and traffic model

The data centers are inherently heterogeneous. Upgrade cycles and replacement of failed components and systems contribute to this heterogeneity. The data centers are expected to upgrade their compute and storage infrastructure to generate different systems that have either (i) different processor architectures, cores and frequencies, (ii) varying memory capacity and interconnection speeds, or (iii) different I/O capabilities. On the other hand, the data centers receive various types of traffic, this traffic having different characteristics and requirements (e.g., voice, video, best effort, etc.). For these reasons, we propose a new system model (similar to the model of the previous section) that would allow the data centers to provide the services needed with lower power consumption.

The model of Section 2.1 addresses the problem of determining the best expansion strategy a cloud provider can take to face the increasing demands and increase its revenue, but it does not take into account the heterogeneity of resources and traffic types. We extend this model by allowing heterogenous set of resources to be available and exploit this heterogeneity in workload distribution by assigning the workloads to the servers customized for the their types. If no such servers are available, we assign the workloads to servers with the goal of minimizing power consumption which would eventually lead to reduced OPEX.

We formulate new design optimization problems using mixed integer-linear programming to address the problem of heterogeneity of resources and traffic that data centers receive. Before describing our model, we briefly go over the notations used and the assumptions made. The new notations used in this model are F to denote the set of all traffic types that data centers receive and R to represent the types of server in the data centers.

The model discussed in Section 2.1 is reformulated as follows.

$$\underset{x,m}{\text{Maximize}} \quad \mathrm{RV}(T) - (\mathrm{OPEX}(T) + \mathrm{CAPEX}(T))$$

$$\text{Subject to} \quad \text{Constraints } (12)\text{—}(21).$$

$$\sum_{s\in S}\lambda_{s,u}^{t,h,f} = L_u^{t,h,f}, \quad \forall f\in F,\ h\in H,\ t\in T \tag{12}$$

$$\sum_{r\in R}\psi_{s,u,f}^{t,h,r} = \lambda_{s,u}^{t,h,f} \quad \forall s\in S,\ f\in F,\ h\in H,\ t\in T \tag{13}$$

$$y_{s,u}^{t,h} \leqslant x_s^t, \quad \forall s\in S,\ u\in U,\ h\in H,\ t\in T \tag{14}$$

$$0 \leqslant \lambda_{s,u}^{t,h,f} \leqslant y_{s,u}^{t,h}L_u^{t,h,f}, \quad \forall s\in S,\ u\in U,\ f\in F,\ h\in H,\ t\in T \tag{15}$$

$$x_s^t M^{\min,r} \leqslant m_s^{t,r} \leqslant x_s^{t,r}M^{\max,r}, \quad \forall r\in R,\ s\in S \tag{16}$$

$$P_{s,f}^{t,h,r} = \left[m_s^{t,r}(P_{\mathrm{idle}} + (E_{\mathrm{usage}}-1)P_{\mathrm{peak}}) + m_s^{t,r}(P_{\mathrm{peak}}-P_{\mathrm{idle}})\gamma_s^{t,h} + x_s^t\epsilon\right]\times \mathrm{MP}_f^r \tag{17}$$

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

$$\gamma_s^{t,h} = \frac{\sum_{u \in U} \psi_{s,u,f}^{t,h,r}}{m_s^{t,r} \mu_s^{t,r}} \tag{18}$$

$$\sum_{f,r \in F,R} P_{s,f}^{t,h,r} \leqslant P_s^{t,h,\max}, \quad \forall s \in S,\ h \in H,\ t \in T \tag{19}$$

$$2 D_{s,u}^{f} y_{s,u}^{t,h} \leqslant D^{f,\max}, \quad \forall s \in S,\ f \in F,\ u \in U,\ h \in H \tag{20}$$

$$\gamma_s^{t,h} \leqslant \gamma^{\max}, \quad \forall s \in S,\ h \in H \tag{21}$$

CAPEX is expressed using Eq. (10), whereas OPEX is expressed using a modified version of Eq. (11) as follows.

$$\text{OPEX(t)} = \sum_{s \in S} \sum_{h \in H} \left(\theta_s^t \sum_{f,r \in F,R} P_{s,f}^{t,h,r} + \delta_s^t (\rho_s + 1) \sum_{f,r \in F,R} P_{s,f}^{t,h,r} + \sum_{u \in U} (\lambda_{s,u}^{t,h,f} \sigma_{s,u}^t) \right)$$

Now, the revenue of year t is computed using the following equation [13]:

$\text{RV}(t) = ((1 - p(x))\alpha^{t,f} \lambda_{s,u}^{t,h} - p(x)\beta^{t,f})$, where $\alpha^{t,f}$ and $\beta^{t,f}$ are the service fee and the penalty for every service request of type f.

A large data center does not have such a completely homogeneous set of servers due to the rapid development of high-performance CPU technologies, data center repairs, replacement, and expansion. Power management for data centers equipped with heterogeneous servers is more complicated due to the challenges associated with how to distribute incoming requests among the different available servers and how to dynamically configure servers to balance the power, performance and reliability trade-off.

In this model, we propose a new distribution request policy to route the parts of the service requests that the data center receives from different user locations to the different servers. For this purpose, let $\psi_{s,u,f}^{t,h,r}$ denote the portion of $\lambda_{s,u}^{t,h,f}$ that is planned to be routed towards server of type r in the data center at location s at each hour h in year t. Let $\Psi = \{\psi_{s,u,f}^{t,h,r} | s \in S, u \in U, f \in F, r \in R, t \in T, h \in H\}$. The constraints of Eqs. (13) and (12) ensure that no request is denied. Eq. (17) calculates the power consumption in a way similar to what is discussed in Section 2.1 except that it uses MP, which represents an appropriate mapping between the traffic types and the available server types. For a certain type of traffic, different server types vary in their suitability to handle this traffic due to the variation in their hardware specifications. So, for a certain service request, the more suitable the server to which it is routed, the less power is consumed in handling it. MP_f^r is used to represent the payoff for assigning request of type f to server of type r. An appropriate mapping of traffic types to server type is recommended. It is possible to determine the minimum number of servers needed in order to support a certain load of traffic and the traffic mix involved. It is hoped that the results will help in understanding how traffic types should be mapped to different server types, and in the definition of appropriate admission control policies.

2.3. 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) "saved" 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 $A = V\left(1 + \frac{i}{n}\right)^{nt}$, where A is the amount of money accumulated after t years, including interest, V is the principal amount (the initial investment amount borrowed or deposited), i is the annual interest rate (as a decimal), t is number of years the amount is deposited or borrowed, and n is the number of times the interest is compounded per year. The Compound Annual Growth Rate (CAGR) is the interest rate at which a given present value would "grow" to a given future value in a given amount of time. It is computed as $\text{CAGR} = \left(\frac{\text{FV}}{\text{PV}}\right)^{\frac{1}{t}} - 1$, where: FV and PV are the future and present values, respectively. Finally, the formula for the inflation rate is $V_d = V(1 + j)^d$, where V_d is total inflated/estimated cost, j is the inflation rate and d is the difference between the base year and the selected year. Alternatively, we can use the following simpler (linear) equation to compute inflation in a much more efficient way $V_d = V + (V \times j \times d)$, which is what we use in our experiments.

2.4. Renewable energy

The model discussed so far does not explicitly account for renewable energy, which is one of the biggest concerns related to data centers and their effect on the surrounding environment. To address this issue, we reformulate Eq. (5) as follows [13].

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

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

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 data center. If $P = G$ then the data center operates as a zero-net energy facility. Finally, if $P < G$, then $[P - G]$ is negative and the power flow is in the direction from the data center to the power grid [13]. Note that, $[P - G]^+$ indicates the amount of power to be purchased from the grid. If this term is negative, the data centers electricity cost will be zero, given the assumption that the grid does not provide compensation for the injected power [13]. In the simulations, we are forced to ignore this important extension due to the lack of realistic input data for the different types of renewable energy generators. Finally, it is worth mentioning that Berral et al. [11] present a different model to handle renewable energy.

3. Simulation results

In this section we present the simulation results of the proposed optimization models discussed in Sections 2.1 and 2.2. The simulation experiments are conducted on a virtual machine running Windows 7 (64-bit) with 16 GB of RAM and 4 processors.

The optimization problems in the two models are solved using CPLEX[2] with Microsoft Visual Studio. CPLEX is a mixed integer-linear programming solver that works based on variations of the branch-and-bound algorithm for integer programming [14] and other metaheuristic methods [15], which can be used to obtain efficient sub-optimal solutions by relaxing the integer constraints. The computational complexity is not a concern in our proposed model as deciding on the best expansion strategy for the cloud provider can be done over offline computations.

A detailed description of the simulation environment for the expansion strategy optimization model is discussed in Section 3.1 while the related simulation results are provided in Section 3.2. Similarly, the simulation of the heterogeneity of resources and traffic model is discussed in Sections 3.3 and 3.4.

3.1. Simulation environment setup for the expansion strategy optimization model

In this section, we discuss the simulation experiments conducted on the expansion strategy optimization model discussed in Section 2.1. 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 are 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 [16], the excluded states are California, Nevada, Idaho, South Dakota, Minnesota, Wisconsin, Ohio, Tennessee, Florida, North Carolina, Virginia, Maryland, New York, Delaware, New Jersey, Connecticut, Rhode Island, Vermont, Massachusetts, and District of Columbia. As for the remaining states, to ensure that Constraint (7) is satisfied, we consider the maximum available power in each state as follows. For Iowa, Kentucky and Mississippi, the maximum available power is 60 MW/h while other states such as Washington, New Hampshire, Oregon, Oklahoma, Utah, Wyoming, Illinois, Arizona, Pennsylvania and South Carolina can handle larger demands (greater then 100 MW/h). The considered candidate locations are depicted in Fig. 2.

After deciding the set S, we turn our attention to other input parameters. According to [10], we set P_{peak} and P_{idle} to 140 and 84 W. A fixed value of 2 is a common choice in the literature for PUE [9]. However, we do consider a more realistic case where the PUE changes with varying outside temperature as shown in Fig. 3. 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 seasons and whether it is during daytime or nighttime. The temperatures are taken from online weather websites such as weatherbase.com and worldweatheronline.com. The details of these periods are shown in Table 1.

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/s [18]. We assume that each server can process one request per second, i.e., $\mu = 3600$. We set $\gamma^{\max} = 0.8$ [19]. The electricity price information based on the average price for industrial load is available at [20]. As mentioned in Section 2.1, 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/kW h. Table 1 shows the details of the considered time periods. Finally, Table 2 shows the values used for the input parameters.

3.2. Simulation results for the expansion strategy optimization model

In this section, we present and discuss the results of the two experiments we conduct to evaluate the expansion strategy optimization model. 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.

[2] http://www.ilog.com/products/cplex/.

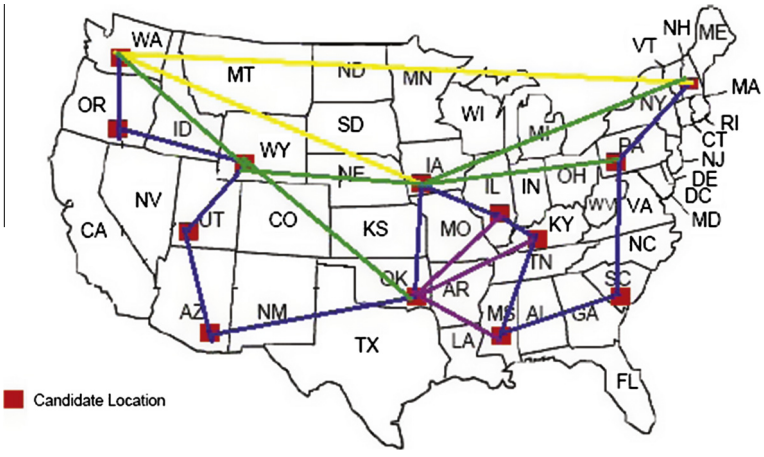

Fig. 2. Candidate locations for data centers in contiguous USA.

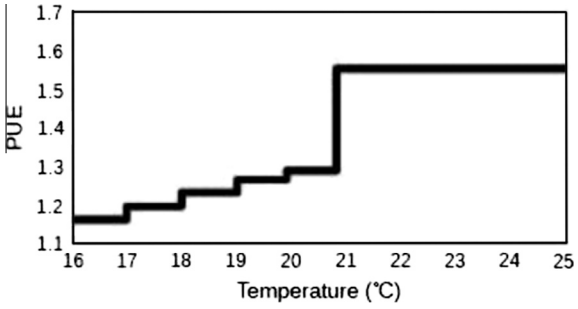

Fig. 3. PUE values for different outside temperature [17].

Table 1
Time periods for both varying PUE values and dynamic pricing.

Period\Season	Summer: May 1–October 31	Winter: November 1–April 30
Daytime	6 am–7 pm	7 am- 6 pm
Nighttime	7 pm–6 am	6 pm–7 am
Off-peak	7 pm–7 am	7 pm–7 am
Mid-peak	7–11 am and 5–7 pm	11 am–5 pm
On-peak	11 am–5 pm	7–11 am and 5–7 pm

Table 2
Input parameters and their values.

Input parameter	Value
P^{max}	60 mW/h (for IA, KY & MS) & 100 mW/h (for WA, NH, OR, OK, UT, WY, IL, AZ, PA & SC)
P_{peak} & P_{idle}	140 W & 84 W
L	[1.5, 2] million hits/s
μ	3600
γ^{max}	0.8
M^{min} & M^{max}	5000 & 50,000

We run the model for five years on 12 data center locations, half of which have existing data centers. Remember that an existing data center 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 data center is an indication of how much traffic load it is processing. We assume that the number of servers to be placed in a single data center ranges between 5000 and 50,000.

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

Table 3
The number of servers in each data centers as computed by our optimization model.

Data center\Year	1	2	3	4	5
DC1	5000	5000	15,122	31,411	46,784
DC2	0	5000	30,237	50,000	50,000
DC3	0	0	0	0	5000
DC4	41,677	47,097	49,999	50,000	50,000
DC5	41,677	50,000	50,000	50,000	50,000
DC6	0	0	0	0	6971
DC7	0	5000	15,122	31,402	50,000
DC8	0	5000	15,123	31,394	50,000
DC9	41,677	50,000	50,000	50,000	50,000
DC10	0	0	0	12,797	50,000
DC11	36,678	47,097	50,000	50,000	50,000
DC12	41,677	50,000	50,000	50,000	50,000
Total	208,386	264,194	325,603	407,004	508,755

In this set of experiments, we consider CAPEX as follows. From Eq. (10), CAPEX depends on two aspects: the building construction cost and the server cost. For the server cost, we assume it is $2000 per server as indicated by [11]. As for the construction cost, we benefit from many websites such as [21] and thecloudcalculator.com to the cost for each of the considered states. We start by estimating the cost of constructing a data center in a reference state (for our experiments, we use NY) to be $20 M. Then, we use several websites such as numbeo.com to compare the cost of living/owing property/building property in NY with the other states of interest. For example, we estimate that building a data center in Oregon costs almost one third of the cost of building it in NY. Finally, we assume that the cloud provider starts with an initial investment of $500 M that will cover CAPEX for the first year. Any construction cost in subsequent years is covered by the leftover as well as the profits generated up until the construction time.

Table 3 shows the number of servers in each data center as computed by our optimization model.[3] The table shows that these numbers increase with the passage of time and the increase in the traffic load. At the beginning, only one (DC1) out of the six existing data centers is lightly loaded small while the other five are heavily loaded. In the second year, the lightly loaded data center (DC1) remains lightly loaded (probably due to its high operational cost or high delay compared with the other available data centers) while the heavily loaded data centers almost reach their full capacity. Moreover, three new data centers are built. The same trend continues in the following years. Data centers 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 data centers are built or the data centers with high operational cost are expanded depending on which option provides higher profits. By the last year of this experiment, the cloud provider is forced to build data centers in all locations to process the huge amount of service requests.

In the second experiment, we study the effect of using a 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:

- Case 1: Fixed PUE and flat rate prices.
- Case 2: Fixed PUE and dynamic prices.
- Case 3: Varying PUE and flat rate prices.
- Case 4: Varying PUE and dynamic prices.

Figs. 4 and 5 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 Fig. 4 shows a significant increase in the actual gained profits of each year, Fig. 5 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 increasing their value compared with non-inflated profits.

From Fig. 4, it can be seen that using dynamic pricing generates (an average of 2%) better annual profits than using fixed pricing. Moreover, using different PUE values for different times of the day 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 Fig. 5 for the annual inflated profits.

3.3. Simulation environment setup for the heterogeneity of resources and traffic model

In this section, we discuss the simulation experiments conducted on the heterogeneity of resources and traffic model discussed in Section 2.2 and the obtained results. As for the simulation setting, we use the same setting defined in Section 3.1 with some modifications (explicitly mentioned) to account for the new parameters of this model as well as to to reduce the

[3] Note that our model computes many parameters including the set M which contains the numbers of servers in each data center for each year.

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

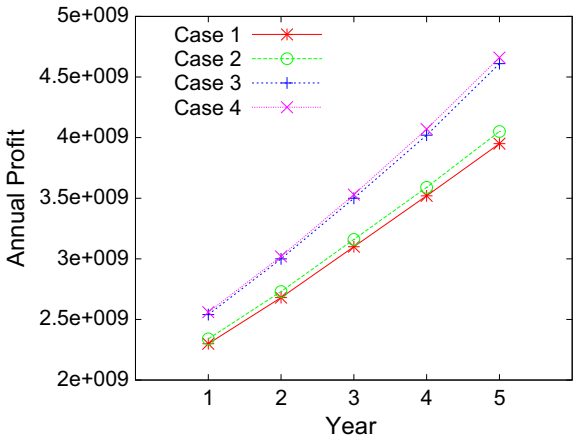

Fig. 4. The original profits for the four cases under consideration.

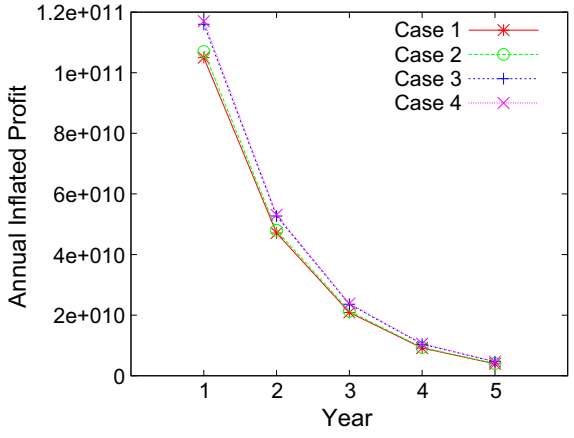

Fig. 5. The inflated profits for the four cases under consideration.

number of variables. For the sake of simplicity, we use fixed values for PUE (1.14 [9]) and electricity prices. As for the traffic load, for the purpose of our study, we assume that there are $F = 4$ different service types, and we choose the total number of service requests for each service type incoming from all user locations to be between 1.5 and 2 million hits/s [18]. We consider four different types of server, $R = 4$, and assume that the number of servers of each type to be placed in a single data center ranges between 2000 and 20,000. Each server has different capacity μ to handle service requests in one second. The details of the processing capacity for each type of server are shown in Table 4. Note that we consider three types of specialized servers (customized for CPU intensive, memory intensive and I/O intensive workloads) in addition to a general purpose type that we call "mixed" server type. Finally, Table 5 shows the values used for the input parameters that are different from Table 2.

3.4. Simulation results for the heterogeneity of resources and traffic model

As with Section 3.2, we conduct two experiments; however, the focus here is on heterogeneity of traffic and servers. In the first experiment, we study the effect of using different types of traffic load to handle different types of servers. We run the

Table 4
The server capacity for each type of service request.

Server type	Processing request per second μ
CPU intensive	5050
Memory intensive	4000
I/O intensive	3600
Mixed	2000

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

Table 5
Input parameters specific to the heterogeneity of resources and traffic model experiments and their values.

Input parameter	Value
M^{min} & M^{max}	2000 & 20,000
E_{usage}	1.4
R	4

model for five years on eight data center locations, three of which have existing data center. Tables 6–9 show how the numbers of servers of each type increase in each data center with the passage of time and the increase in the traffic load.

In the second experiment, we study the effect of handling different types of traffic load using a single type of servers (i.e., using a homogenous set of servers). We assume that each server can process one request per second, i.e., $m = 3600$.

Figs. 6 and 7 show the annual profits (original and inflated) generated for the two cases under consideration The figures show that handling different types of traffic loads using different types of servers generates an average of 5% better annual profits than when using a single type of servers.

4. Literature review

The main problem addressed in this work is the expansion strategies of cloud providers to meet the increasing user demands. The body of work on this problem is limited since most of the current works focus on optimizing the currently

Table 6
The number of CPU intensive servers in each data center.

Data center\Year	1	2	3	4	5
DC1	0	0	0	0	0
DC2	0	0	0	0	4750
DC3	0	0	0	2000	20,000
DC4	0	0	2000	20,000	20,000
DC5	0	2000	20,000	20,000	20,000
DC6	6384	20,000	20,000	20,000	20,000
DC7	20,000	20,000	20,000	20,000	20,000
DC8	20,000	20,000	20,000	20,000	20,000
Total	44,217	62,000	90,486	98,665	108,000

Table 7
The number of memory intensive servers in each data center.

Data center\Year	1	2	3	4	5
DC1	0	0	0	0	0
DC2	0	0	0	0	20,000
DC3	0	0	0	2000	20,000
DC4	0	0	2001	20,000	20,000
DC5	0	14,480	20,000	20,000	20,000
DC6	20,000	20,000	20,000	20,000	20,000
DC7	20,000	20,000	20,000	20,000	20,000
DC8	20,000	20,000	20,000	20,000	20,000
Total	24,000	31,263	46,000	48,001	70,000

Table 8
The number of I/O intensive servers in each data center.

Data center\Year	1	2	3	4	5
DC1	0	0	0	0	0
DC2	0	0	0	0	20,000
DC3	0	0	0	2000	20,000
DC4	0	0	2000	20,000	20,000
DC5	0	2000	20,000	20,000	20,000
DC6	20,000	20,000	20,000	20,000	20,000
DC7	20,000	20,000	20,000	20,000	20,000
DC8	20,000	20,000	20,000	20,000	20,000
Total	44,217	62,000	90,486	98,665	108,000

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

Table 9
The number of mixed servers in each data centers.

Data center\Year	1	2	3	4	5
DC1	0	0	0	0	0
DC2	0	0	0	0	2000
DC3	0	0	0	2000	2000
DC4	0	0	2000	19,000	20,000
DC5	0	2000	17,600	20,000	20,000
DC6	2000	19,999	20,000	20,000	20,000
DC7	20,000	20,000	20,000	20,000	20,000
DC8	20,000	20,000	20,000	20,000	20,000
Total	24,000	31,263	46,000	48,001	70,000

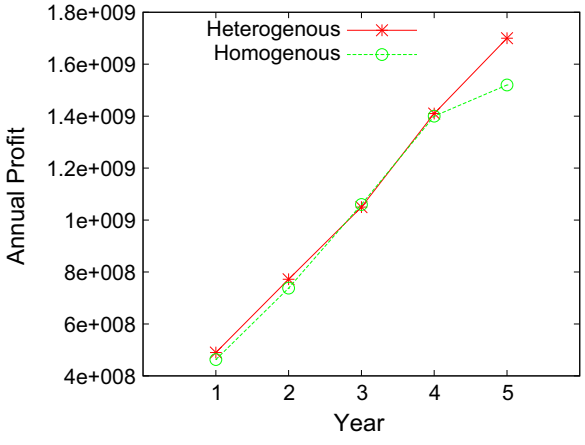

Fig. 6. The original profits for the two experiments under consideration.

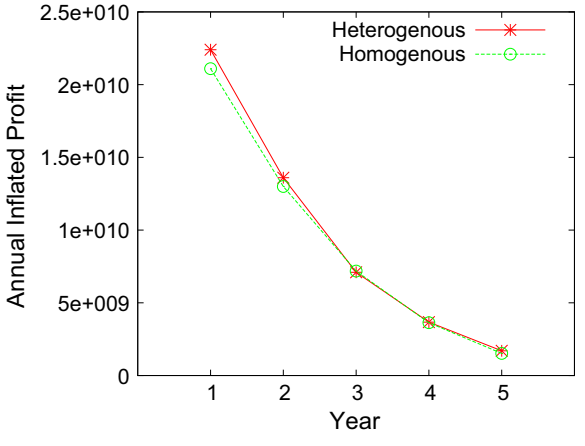

Fig. 7. The inflated profits for the two experiments under consideration.

available data centers by improving power consumption, cooling, request routing, etc. For a broader coverage of such issues, the interested readers are referred to recent surveys such as [22]. We start our discussion of the related work by discussing these issues before going into the more relevant papers concerned with the added issue of building new data centers and/or expanding existing data centers by increasing their service capacity, which is achieved by increasing the number of servers they contain.

A rich volume of recent research work focused on reducing power costs instead of consumptions. These research works mainly devise different workload distribution policies across geo-distributed data centers for achieving different performance objectives such as total electricity cost minimization [18,23–30], bandwidth cost minimization [31], energy efficiency

improvement [32–35], cooling efficiency [36], carbon footprint minimization [37], and renewable energy usage maximization [38–40]. In order to achieve these objectives, researchers mainly formulated the workload distribution problem as various linear and non-linear optimization problems and adopted various methods and tools to solve them. For optimum solutions, the commonly used mathematical tools are mixed integer programming [23,25,26,41].

Qureshi et al. [18] is one of the first works to focus on the temporal and spatial variability of electricity prices in the wholesale market. The authors argued that since the electricity prices fluctuate across different regions energy expense per unit of computation is not the same for every data center. 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 data center located at cheaper price regions within some radial geographical distance.

The recent research focused on the future site of the data center, because the electricity price is not the same in each region. The data center is expected to be built in locations with lower prices, colder weather and available renewable energy to reduce the carbon footprint.

In [8], the authors studied the problem of selecting the best locations to build a fixed number of data centers. They assumed that the data centers to be built are not associated with any already existing data centers. They formulated three optimization problems with three objective functions. The objective functions are (i) minimizing the carbon footprint, (ii) minimizing the total cost (including energy cost, bandwidth cost and carbon tax), and (iii) minimizing the average service latency subject to the QoS constraints. Their formulation takes as input a set of candidate locations and determines the best location(s) to build future data center(s), the number of servers required at each new data center, and how the service requests can be routed to each data center.

Another work to select future site(s) is proposed in [17]. The authors introduced a new process that can be used to select the best location to build new data centers while considering any already existing data centers the cloud provider may have. The authors formulated an optimization problem considering the following factors: (i) the capital cost (CAPEX), which includes the costs of land acquisition, construction of the data center's infrastructure, electricity and bandwidth supplied to the data center, etc., (ii) the operational cost (OPEX), which includes the costs of electricity, bandwidth, water for cooling the data center, carbon tax, administration staff salary, etc., (iii) response time, which depends on the distance between the location of data center and a population center, (iv) consistency delay,[4] which depends on the distance between two potential locations of the data center 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. In a later extension of this work, Berral et al. [11] presented a new formulation with similar objective but with more involved and complex constraints that focus more on renewable energy.

Reducing the carbon footprint and maximizing renewable energy usage are the objectives considered by the authors of [42]. They showed that the carbon footprint can be reduced by building the data centers near the sites of renewable energy. They developed a mathematical model for calculating the total carbon footprint including: manufacturing, usage and communication footprints. They discussed how to reduce the carbon footprint (especially, the manufacturing footprint) by redistributing the load of the data center to other sites based on the availability of renewable energy.

The authors of [23] 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 considered total electricity cost, load constraint, end-to-end delay constraints for data centers.[5] 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 [43].

Compared with the previously mentioned related works, the contributions of this work lie in the following points. To the best of our knowledge, no prior work has addressed the problem of determining the best future location of the data center while taking into account the tradeoff between maximizing the revenue and minimizing the operational cost of the data center instead of (bandwidth, cooling, carbon tax, and power costs). Moreover, previous works neglected important economical aspects such as the annual inflation in the costs (bandwidth, cooling, price of electricity) and in the revenue. Finally, our proposed model takes into account heterogeneity in both traffic as well as resources, which is more realistic than the commonly assumed homogeneity in traffic and resources.

5. Conclusion and future work

With the growth of ultrascale data centers around the world, research on reducing operational cost (OPEX) in the data center is still in its infancy. 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 data centers or to simply expand the data centers it currently has. Choosing future sites for constructing new data centers requires careful consideration on several factors, reducing the electricity cost, bandwidth cost, carbon footprint, maximizing renewable energy usage, and QoS, etc. The data centers are inherently heterogeneous, the data centers receive various types of traffic with different characteristics

[4] When a data center becomes unreachable or unavailable, the data centers that provide a service must be mirrors of each other. The consistency delay refers to the time required for state changes to reach all mirrors.

[5] The queuing delay inside the data center is assumed to be an M/M/1 queue.

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

and requirements. We proposed a formulation of the problem that takes into account the locations and capacities of the future data centers, the operational cost of the data centers, the heterogeneity of resources and traffic that should satisfy some delay limitations, and important economical aspects such as the annual inflation in the costs and revenue.

Few major lessons that we have found from this study are as follows. First, minimizing electricity cost causes workload to follow the cheaper price regions which does not necessarily minimize carbon footprint. Second, renewable energy sources are intermittent and unpredictable; data centers can leverage effective integration of such renewable. Finally, traffic and computing resources heterogeneity could increase the room for better optimizing the usage of today's data centers.

Our work can be extended in many directions to reduce the energy cost and power density in data centers. One direction is to consider the server consolidation, which refers to assigning incoming tasks to the minimum number of active servers in the data center and shutting down unused servers. Thus, it would be interesting to extend the revenue maximization problems introduced in the previous sections by taking into account the cost for running the servers and assuming that the machines can be switched on and off dynamically. We will be considering this as a future work.

References

[1] J. Shi, M. Taifi, A. Khreishah, Resource planning for parallel processing in the cloud, in: 2011 IEEE 13th International Conference on High Performance Computing and Communications (HPCC), IEEE, 2011, pp. 828–833.
[2] J. Shi, M. Taifi, A. Khreishah, J. Wu, Sustainable GPU computing at scale, in: 2011 IEEE 14th International Conference on Computational Science and Engineering (CSE), IEEE, 2011, pp. 263–272.
[3] J. Koomey, Growth in Data Center Electricity Use 2005 to 2010, Analytics Press, Oakland, CA, 2011.
[4] A. Rahman, X. Liu, F. Kong, A survey on geographic load balancing based data center power management in the smart grid environment, Commun. Surv. Tutorials, IEEE 16 (1) (2014) 214–233.
[5] Y. Jararweh, A. Hary, Y. Al-Nashif, S. Hariri, A. Akoglu, D. Jenerette, Accelerated discovery through integration of kepler with data turbine for ecosystem research, in: Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on, 2009, pp. 1005–1012, http://dx.doi.org/10.1109/AICCSA.2009.5069454.
[6] Y. Jararweh, S. Hariri, Power and performance management of GPUs based cluster, Int. J. Cloud Appl. Comput. 2 (4) (2012) 16–31, http://dx.doi.org/10.4018/ijcac.2012100102.
[7] M. Wardat, M. Al-Ayyoub, Y. Jararweh, A.A. Khreishah, To build or not to build? Addressing the expansion strategies of cloud providers, in: 2014 International Conference on Future Internet of Things and Cloud (FiCloud), IEEE, 2014, pp. 477–482.
[8] A.-H. Mohsenian-Rad, A. Leon-Garcia, Energy-information transmission tradeoff in green cloud computing, Carbon 100 (2010) 200.
[9] R. Brown et al., Report to Congress on Server and Data Center Energy Efficiency: Public Law 109-431, 2008.
[10] X. Fan, W.-D. Weber, L.A. Barroso, Power provisioning for a warehouse-sized computer, SIGARCH Comput. Archit. News 35 (2) (2007) 13–23, http://dx.doi.org/10.1145/1273440.1250665.
[11] J.L. Berral, Í. Goiri, T.D. Nguyen, R. Gavalda, J. Torres, R. Bianchini, Building green cloud services at low cost, in: 2014 IEEE 34th International Conference on Distributed Computing Systems (ICDCS), IEEE, 2014, pp. 449–460.
[12] S. Baldwin, Carbon footprint of electricity generation, London: Parliament. Office Sci. Technol. 268 (2006). <http://www.parliament.uk/documents/post/postpn268.pdf>.
[13] M. Ghamkhari, H. Mohsenian-Rad, Energy and performance management of green data centers: a profit maximization approach, IEEE Trans. Smart Grid 4 (2) (2013) 1017–1025.
[14] D.P. Bertsekas, Nonlinear Programming, Athena Scientific Belmont, 1999.
[15] H.R. Lourenço, O.C. Martin, T. Stützle, Iterated Local Search, Springer, 2003.
[16] Energy Information Administration, State Electricity Profiles – Summer Capacity, 2008.
[17] Í. Goiri, K. Le, J. Guitart, J. Torres, R. Bianchini, Intelligent placement of datacenters for internet services, in: 2011 31st International Conference on Distributed Computing Systems (ICDCS), IEEE, 2011, pp. 131–142.
[18] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, B. Maggs, Cutting the electric bill for internet-scale systems, ACM SIGCOMM Comput. Commun. Rev. 39 (4) (2009) 123–134.
[19] K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiran, F. Tobagi, C. Diot, Analysis of measured single-hop delay from an operational backbone network, INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 2, IEEE, 2002, pp. 535–544.
[20] Energy Information Administration, Average Retail Price of Electricity to Ultimate Customers by End-use Sector by State, 2010.
[21] Emerson Network Power, Higher Density = Lower Cost, Efficiency Without Compromise E-Book Series, 2009.
[22] J. Shuja, K. Bilal, S. Madani, M. Othman, R. Ranjan, P. Balaji, S. Khan, Survey of techniques and architectures for designing energy-efficient data centers, Syst. J., IEEE PP (99) (2014) 1–13, http://dx.doi.org/10.1109/JSYST.2014.2315823.
[23] L. Rao, X. Liu, L. Xie, W. Liu, Minimizing electricity cost: optimization of distributed internet data centers in a multi-electricity-market environment, in: INFOCOM, 2010 Proceedings IEEE, IEEE, 2010, pp. 1–9.
[24] L. Rao, X. Liu, M. Ilic, J. Liu, Mec-idc: joint load balancing and power control for distributed internet data centers, in: Proceedings of the 1st ACM/IEEE International Conference on Cyber-Physical Systems, ACM, 2010, pp. 188–197.
[25] L. Rao, X. Liu, L. Xie, W. Liu, Coordinated energy cost management of distributed internet data centers in smart grid, IEEE Trans. Smart Grid 3 (1) (2012) 50–58.
[26] Y. Zhang, Y. Wang, X. Wang, Capping the electricity cost of cloud-scale data centers with impacts on power markets, in: Proceedings of the 20th International Symposium on High Performance Distributed Computing, ACM, 2011, pp. 271–272.
[27] Y. Yao, L. Huang, A. Sharma, L. Golubchik, M. Neely, Data centers power reduction: a two time scale approach for delay tolerant workloads, in: INFOCOM, 2012 Proceedings IEEE, IEEE, 2012, pp. 1431–1439.
[28] Y. Guo, Z. Ding, Y. Fang, D. Wu, Cutting down electricity cost in internet data centers by using energy storage, in: Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE, IEEE, 2011, pp. 1–5.
[29] P. Ostovari, J. Wu, A. Khreishah, The benefits of cooperation between the cloud and private data centers for multi-rate video streaming, in: 23rd International Conference on Computer Communications and Networks (ICCCN 2014), 2014.
[30] M. Taifi, J.Y. Shi, A. Khreishah, SpotMPI: a framework for auction-based HPC computing using amazon spot instances, in: Algorithms and Architectures for Parallel Processing, Springer, 2011, pp. 109–120.
[31] N. Buchbinder, N. Jain, I. Menache, Online job-migration for reducing the electricity bill in the cloud, in: NETWORKING 2011, Springer, 2011, pp. 172–185.
[32] A.N. Sankaranarayanan, S. Sharangi, A. Fedorova, Global cost diversity aware dispatch algorithm for heterogeneous data centers, SIGSOFT Softw. Eng. Notes 36 (5) (2011) 289–294, http://dx.doi.org/10.1145/1958746.1958787.
[33] J. Li, Z. Li, K. Ren, X. Liu, Towards optimal electric demand management for internet data centers, IEEE Trans. Smart Grid 3 (1) (2012) 183–192.

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002

[34] L. Rao, X. Liu, M.D. Ilic, J. Liu, Distributed coordination of internet data centers under multiregional electricity markets, Proc. IEEE 100 (1) (2012) 269–282.
[35] L. Parolini, B. Sinopoli, B.H. Krogh, Z. Wang, A cyber-physical systems approach to data center modeling and control for energy efficiency, Proc. IEEE 100 (1) (2012) 254–268.
[36] K. Le, R. Bianchini, J. Zhang, Y. Jaluria, J. Meng, T.D. Nguyen, Reducing electricity cost through virtual machine placement in high performance computing clouds, in: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, ACM, 2011, p. 22.
[37] J. Doyle, D. O'Mahony, R. Shorten, Server selection for carbon emission control, in: Proceedings of the 2nd ACM SIGCOMM Workshop on Green Networking, ACM, 2011, pp. 1–6.
[38] A. Krioukov, C. Goebel, S. Alspaugh, Y. Chen, D.E. Culler, R.H. Katz, Integrating renewable energy using data analytics systems: challenges and opportunities, IEEE Data Eng. Bull. 34 (1) (2011) 3–11.
[39] R. Carroll, S. Balasubramaniam, D. Botvich, W. Donnelly, Dynamic optimization solution for green service migration in data centres, in: 2011 IEEE International Conference on Communications (ICC), IEEE, 2011, pp. 1–6.
[40] Y. Zhang, Y. Wang, X. Wang, Greenware: greening cloud-scale data centers to maximize the use of renewable energy, in: Middleware 2011, Springer, 2011, pp. 143–164.
[41] Z. Abbasi, T. Mukherjee, G. Varsamopoulos, S.K. Gupta, Dynamic hosting management of web based applications over clouds, in: 2011 18th International Conference on High Performance Computing (HiPC), IEEE, 2011, pp. 1–10.
[42] W. Van Heddeghem, W. Vereecken, D. Colle, M. Pickavet, P. Demeester, Distributed computing for carbon footprint reduction by exploiting low-footprint energy availability, Future Gener. Comput. Syst. 28 (2) (2012) 405–414.
[43] U. Brenner, A faster polynomial algorithm for the unbalanced hitchcock transportation problem, Oper. Res. Lett. 36 (4) (2008) 408–413.

Please cite this article in press as: M. Al-Ayyoub et al., Optimizing expansion strategies for ultrascale cloud computing data centers, Simulat. Modell. Pract. Theory (2015), http://dx.doi.org/10.1016/j.simpat.2015.03.002