Simulation Modelling Practice and Theory 58 (2015) 15–29

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 26 March 2015

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.