Mathematical Programming for Server Consolidation in Cloud Data Centers

Bo Wang*, Ying Song†, Xiao Cui*, and Jie Cao*
*Software Engineering College, Zhengzhou University of Light Industry, Zhenzhou, China, 450002
Email: wangb@zzuli.edu.cn, cuixiao_0217@sina.com, cjjiecao@163.com
†Computer School, Beijing Information Science and Technology University
Email: songying@bistu.edu.cn

***Abstract*—Server consolidation based on virtualization technology simplifies system administration and improves energy efficiency by improving resource utilizations and reducing the physical machine (PM) number in contemporary service-oriented data centers. These benefits prompt service providers to deliver their services on virtualized data centers. For a service provider, their total costs are mainly composed of the investment costs for buying infrastructures, such as the ownership costs of PMs, and the operational costs, such as the electricity costs for powering PMs, cooling, lighting and so on. Plenty work has studied on minimizing the operational costs. On the contrary, we study on the scale planning for minimizing the investment costs for building/updating data centers providing Internet services, in this paper. We model the scale planning as an integer program minimizing the total ownership costs of PMs. Extensive experiments results show that our scale planning is better than the plannings made by static and dynamic consolidations and that using the scale planning improves the computing times consumed by online consolidations without impact on their performance.**

I. Introduction

Cloud computing and data centers have become an important part of our daily lives because of various Internet services, such as Internet-wide search and e-mail services. The scales of cloud data centers in the number of physical machines (PMs) increase with Internet users and their demands for Internet resources. Thousand-node data centers are increasingly common. While, most of the time the resource utilization of a data center is below 50% in real world [1], [2], which leads to many useless costs for service providers. This motivates many researchers studying on improving these useless costs.

For data center operators, their total costs are mainly composed of two parts: [3].

- the **investment costs** for buying infrastructures including PMs, network devices, cooling equipments, lighting equipments and other auxiliaries when a data center operator wants to build a new data center or switch to a new technology of hardware, such as a new generation of multicore processors. These costs could not be reduced by runtime managements.
- the **operational costs** which consist of the electricity costs for power, such as powering PMs, cooling, lighting and so on, software copyright costs, hardware/software maintenance costs, and so on. The works focusing on these costs consider investment costs as "sunk costs". These costs can be improved by online resource managements.

There have been plenty works [3]–[21] studying on minimizing one or more kinds of operational costs while few work studied on improving investment costs. The Green Grid consortium surveyed 188 data centers in 2010, mostly located in the United States, and estimated that, on average, 10% of servers are never utilised [22]. The costs for these servers can be saved at the beginning of building data centers. Thus, in this paper, we study on minimizing the kind of investment costs, the total cost of ownership for PMs, when service providers plan to update infrastructures of their cloud data centers.

For operating cloud data centers providing elastic Internet services, one of the most challenge for optimizing investment costs is that there are three dimensions should be considered. (i) The first one is the variations of various resource requirements for services over time, and the resource competition or complementarity between/amongst services. Consolidating services whose requirements of resources are complementary helps to reduce idle resources for PMs, and vice versa. (ii) The second dimension should be considered is the instance numbers of services. If the instance number is too large for a service, there would be many underutilized VMs when the load is low, which will increase the used PM number, leading to energy inefficiency. While if the number is too small, the performance requirement of the service will not be satisfied when the load is high. (iii) The last one is the resource configurations for each service instance. There will be idle resources if an instance is over provision, while a performance violation if under provision.

To address the challenge, in this paper, we present an analysis model to minimizing the total ownership cost of PMs for building/updating date centers providing Internet services. The inputs of this model comprises resource demands of provided Internet services and the ownership costs and resource capacities of PMs to be chose. All of these information can be easily achieved using the logged data from previously running of data centers [3] or theoretical model, such as queuing theory [23], based on the service performance that data center operators want to provide. The resource demands change with time and usually have seasonal patterns on a daily, weekly, or monthly basis [3], [24], [25] in real word. The outputs of our model consist of the minimal cost, the chose PMs and the services deployment at each time. After this, we use a real world trace to quantify the proposed model, and simulation results show our model has superior performance.

The rest of this paper is organized as follows. Section II discusses related work. Section III introduces our mathematical programming model. Section IV presents the experimental results. We conclude this paper in Section V.

978-1-5386-1107-4/17/$31.00 ©2017 IEEE

II. Related work

Many efforts [3], [7]–[13] have examined energy management strategies in data centers. These efforts tackled the high "base" power of traditional server hardware (i.e. the power consumption when the system is powered on but idle) [2], by static or dynamic consolidation.

The static consolidation methods [3], [7]–[9] gave a mapping plan of VMs and physical servers with the input of the maximum performance required by services, which minimizes the number of servers or the overall operational costs at a time interval. These methods had to know the VMs status, i.e. the number of VMs, and the configuration and the profiling data of every VM, which do not need by our model. These methods did not take advantage of the various changes of performance requirements for different services with time and thus would provide the scale larger than optimum if using these methods for scale planing because it is unlike that all services need their maximum performance requirements at the same time at runtime.

The changeable status of VMs offered new opportunities to re-consolidate the running VMs. Dynamic consolidation methods [8]–[13] dynamically reconfigured (or shrinked) a data center to operate with fewer nodes by VM migration. These works were reactive, which made decisions during the process of running the services. While VM migration leads to performance loss and energy overhead that cannot be ignored [26]–[28], and only such dynamic control of turning on/off PMs is not enough to guide the management of data centers for the administrators and designers.

The above methods all aimed at minimizing the energy consumed by computing. While the cooling cost increases with the highest temperature of clusters because of the load of nodes increased by server consolidations. The cooling cost has been about half of the total electricity costs [15]. Therefore, a few works [15]–[21] studied on improving cooling costs. They scheduled tasks/VMs to balance the heat in a data center to decrease the highest temperature (hotspot). The cooling cost was improved by increasing the temperature of cooling air because of the decreased hotspot for maintaining a level of temperatures in the data center, with the basis of cooling energy proportional to temperature of supplied cooling air [15].

All of these above works were improving operational costs, which is orthogonal to our work. Our work is planning the scale of virtualized data centers to minimize investment costs and complements these previous efforts very well. The combination of these reactive works and our work contribute to the wide use of virtualized server consolidation in data centers.

A similar work was proposed by B. Speitkamp and M. Bichler [3]. They proposed a mathematical programming approach to minimize the total cost of ownership for PMs. While their work planed for the services of which each corresponds to exactly one VM. These services' performance can be tuned only by reallocating resources to their corresponding VMs. Our work addresses the problem for Internet services each of which corresponds to multiply VMs. The performance of an Internet service can be tuned by not only reallocating resources to its corresponding VMs but also tuning the number of the VMs.

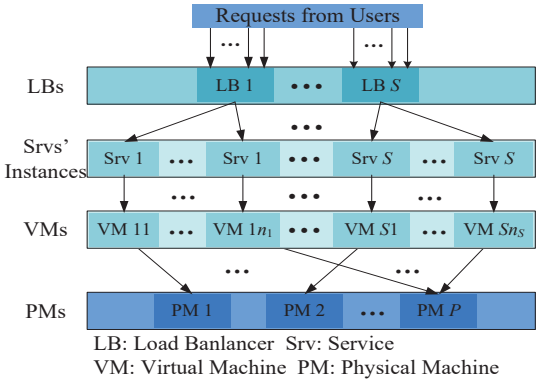

Fig. 1. The architecture of a virtualized data center providing Internet services.

III. Mathematical Programming Model

A. Background

In a virtualized data center providing Internet services, as shown in Fig. 1, a request is distributed to an instance of the corresponding service which has multiple instances each of which is deployed in a VM hosted on a PM, by the corresponding load banlancer (LB). The designs of LBs are out of scope of this paper. In this paper, we focus on planing the scale of data centers, minimizing the ownership costs of PMs with the capacity of providing required performances for services.

B. Planning model

For building/updating a data center, there are P PMs to be chose. The ownership cost of PM k ($k = 1, ..., P$) is c_k. There are R types of resources. The amount of resource j ($j = 1, ..., R$) on PM k is $r_{j,k}$. The data center will be used to provide S services. The required performance of service i ($i = 1, ..., S$) is $\mu_{i,t}$ at time interval t ($t = 1, ..., T$). In this paper, we consider throughput as the performance metric, while using other metrics does not fundamentally alter our approach. At runtime, there are V VM configurations can be chose. For resource j, the configured amount ($l = 1, ..., V$) is $v_{j,l}$ in VM configuration l. On PM k, the performance provided by a VM with configuration l for service i is $\mu_{i,k,l}$. We define the variables $x_{i,k,l,t}$, $i = 1, ..., S$, $k = 1, ..., P$, $l = 1, ..., V$, $t = 1, ..., T$, where $x_{i,k,l,t} = m$ if there are m VMs hosted on PM k with configuration l to provide service i at time interval t, and the binary variables z_k, $k = 1, ..., P$, where $z_k = 1$ if PM k is used at any interval and $z_k = 0$ if not.

We formulate the problem of building the data center with minimal costs as an integer program as follows:

$$\text{Minimize} \sum_{k=1}^{P} (c_k \cdot z_k), \tag{1}$$