Future Generation Computer Systems 48 (2015) 82–95

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Stochastic modeling of dynamic right-sizing for energy-efficiency in cloud data centers

Dian Shen[a,*], Junzhou Luo[a], Fang Dong[a], Xiang Fei[b], Wei Wang[a], Guoqing Jin[b], Weidong Li[b]

[a] *School of Computer Science and Engineering, Southeast University, Nanjing 211189, PR China*

[b] *Faculty of Engineering and Computing, Coventry University, Coventry, United Kingdom*

HIGHLIGHTS

- A Stochastic Right-sizing Model(SRM) based on Queueing theory is proposed.
- A BFGS based algorithm is proposed to achieve energy-efficiency.
- SRM is implemented with open-source cloud platform OpenStack.

ARTICLE INFO

Article history:
Received 4 March 2014
Received in revised form
10 July 2014
Accepted 19 September 2014
Available online 5 October 2014

Keywords:
Cloud computing
Data center
Virtualization
Energy-efficiency
Queueing theory

ABSTRACT

Large data centers are usually built to support increasing computational and data storage demand of growing global business and industry, which consume an enormous amount of energy, at a huge cost to both business and the environment. However, much of that energy is wasted to maintain excess service capacity during periods of low load. In this paper, we investigate the problem of "right-sizing" data center for energy-efficiency through virtualization which allows consolidation of workloads into smaller number of servers while dynamically powering off the idle ones. In view of the dynamic nature of data centers, we propose a stochastic model based on Queueing theory to capture the main characteristics. Solving this model, we notice that there exists a tradeoff between the energy consumption and performance. We hereby develop a BFGS based algorithm to optimize the tradeoff by searching for the optimal system parameter values for the data center operators to "right-size" the data centers. We implement our Stochastic Right-sizing Model (SRM) and deploy it in the real-world cloud data center. Experiments with two real-world workload traces show that SRM can significantly reduce the energy consumption while maintaining high performance.

1. Introduction

Large data centers are usually built to support increasing computational and data storage demand of growing global business and industry. Take Facebook as an example, the number of computation nodes in the data center is up to 60 000. Such large scale data centers generally consume an enormous amount of electrical power resulting in high operational costs and carbon dioxide emissions [1–4]. Currently, data centers that power Internet-scale applications consume about 1.3% of the worldwide electricity supply and this fraction is expected to grow to 8% by 2020 [2]. Moreover, the energy consumption of such data centers all over the world is expected to double in every five years, at a huge cost to both business and the environment. Unfortunately, most of that energy is wasted to maintain excess service capacity during periods of low load. Data collected from more than 5000 production servers over a six-month period have shown that most of the time servers operate at 10%–50% of their full capacity [5], leading to wasting expenses on low utilization of resources.

Therefore, it is of great significance and challenge to manage the capacity of data centers to use power only in proportion to the load, i.e., to "right-size" [6] the data centers. The criteria of right-sizing differ among different data center operators. For the data center operators, they are willing to maximize their benefits through reducing the energy consumption and meet some performance

[*] Corresponding author. Tel.: +86 18652070734.
E-mail addresses: dianshen@seu.edu.cn (D. Shen), jluo@seu.edu.cn (J. Luo), fdong@seu.edu.cn (F. Dong), aa5861@coventry.ac.uk (X. Fei), wangweing@seu.edu.cn (W. Wang), ab4395@coventry.ac.uk (G. Jin), aa3719@coventry.ac.uk (W. Li).

http://dx.doi.org/10.1016/j.future.2014.09.012
0167-739X/© 2014 Elsevier B.V. All rights reserved.

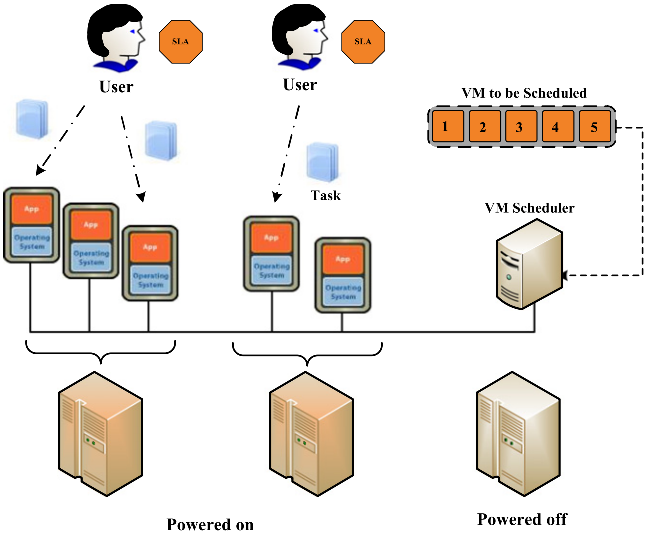

Fig. 1. Architecture of a cloud data center using virtualization technology.

standards (e.g. Service Level Agreements) so as to utilize the resources in a cost-optimal manner, i.e., to achieve energy-efficiency [3]. With the emergence of Cloud Computing [7], virtualization becomes one of the promising technologies to implement energy-efficient right-sizing of data centers, which allows flexibly in the provisioning and placement of servers and workloads in the data centers. The typical architecture of cloud data centers supported by virtualization technologies is illustrated in Fig. 1.

As is shown in Fig. 1, the lowest layer of this architecture is a pool of physical servers organized as the data center infrastructure consisting of large amount of computing and storage resources. Above this layer is the virtualization layer where users access the resources in the form of virtual machines (VMs) and submit their tasks with different Service-Level Agreements (SLAs). From this architecture, it can be seen that with the benefits of VMs, the goal of "right-sizing" the data centers can be accomplished through the consolidation of computation load into smaller number of servers while powering off the idle ones [8–10]. Driven by this idea, the basic problem underlying the right-sizing can be seen as to adapt the number of active servers in the data centers to match the current workload. This problem becomes more challenging when recognizing the dynamic nature of data centers that VMs arrive at the system from time to time and release the resources when finishing their tasks. Therefore, with different VMs arriving and leaving at each moment, the cloud system can be viewed as a highly dynamic system. In view of its dynamic nature, Queueing theory and stochastic process are especially useful in this scenario, which allow us to understand the behavior of such a complex system and to guide the design of data center resource management policies. The main contributions of this paper are summarized as:

(1) In view of the dynamic nature of cloud data centers, we introduce a stochastic model based on Queueing theory which captures the main characteristics in "right-sizing" the cloud data center. We then solve the proposed stochastic model and derive the stationary system state distribution with which the analytical expressions of several important performance measures of the system are given. The model is further extended to adapt to real-world environment.

(2) While improving performance points to turning on more servers, minimizing energy consumption aims at turning on fewer servers, there is clearly a tradeoff between energy consumption and execution time. We hereby develop a Broyden–Fletcher–Goldfarb–Shanno (BFGS [11]) based algorithm to optimize the tradeoff by searching for the optimal system parameter values to achieve a given optimization goal. In the optimization process, we recognize the performance overhead introduced by the virtualization and take it into consideration, generating more accurate optimization results.

(3) We implement the Stochastic Right-sizing Model (SRM) with an open-source cloud platform and deploy it in the real-world cloud data center SEUCloud. Experiments with two real-world workload traces highlight that the cost and energy savings achievable by SRM are significant, while the performance remains high and stable compared with the state-of-art method.

Table 1
A summary of notation for reference.

Symbol	Definition
N	The number of servers in the data center
m	The maximum number of VMs that can be allocated to one server
M	The maximum number of running VMs in the system
λ	The arrive rate of VMs
μ	The service rate of servers
θ	The failure rate of servers
σ	The setup rate of servers
β	The over-provisioning factor
π_{ij}	The probability that there are j VMs and i active servers in the system
Q	The Markov transition matrix
$W_i(t)$	The conditional waiting probability that a VM will begin service no later than time t
$\mathcal{R}$	A specific VM scheduler
C_v	The virtualization overhead
C_h	The cost of holing one VM in the system
C_μ	The cost of one server at service rate μ
C_S	The constant energy cost of one server at *setup* state
$E(V)$	The expected number of VMs in the system
$E(B)$	The expected number of servers at *on* state
$E(S)$	The expected number of servers at *setup* state
$E(O)$	The expected number of servers at *off* state
$E(D)$	The expected number of servers at *fail* state
$E(W)$	The expected sum of VMs' execution time
$E(P)$	The expected sum of energy consumption

The rest of this paper is organized as follows. In Section 2, the stochastic model of data centers is formalized and solved. In Section 3, we analyze the factors that may affect the performance and propose a BFGS based optimization algorithm. Section 4 discusses the extension of our model to adapt into real-world environments. Section 5 describes the design and implementation of our stochastic right-sizing model. Section 6 reports the experiment results with two real-world workload traces. In Section 7, we list the related work. Finally, we have the conclusion on our model in Section 8.

2. Mathematical model

In view of the dynamic nature of cloud data centers, we first introduce a stochastic model based on Queueing theory which captures the main characteristics in the cloud data center environment. We then try to solve the model and derive the stationary system state distribution. We summarize the notation used in the paper in Table 1.

2.1. General model

We consider a cloud data center environment, consisting of N identical servers, and each server can run up to m specific virtual machines. Therefore, the system can have at most $M = m * N$ virtual machines running concurrently; other VMs have to wait in the queue for scheduling. In order to facilitate the analysis of this scenario, we first make a general assumption that VMs arrive at the system according to a Poisson input stream of rate λ. The service