Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 365–374

Procedia Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

A Proactive Cloud Scaling Model Based on Fuzzy Time Series and SLA Awareness

Dang Tran[1], Nhuan Tran[1], Giang Nguyen[2], and Binh Minh Nguyen[1]*

[1] School of Information and Communication Technology,
Hanoi University of Science and Technology, Vietnam
dangtv18@gmail.com, tranducnhuan1994@gmail.com, minhnb@soict.hust.edu.vn
[2] Institute of Informatics, Slovak Academy of Sciences, Slovakia
giang.ui@savba.sk

Abstract

Cloud computing has emerged as an optimal option for almost all computational problems today. Using cloud services, customers and providers come to terms of usage conditions defined in Service Agreement Layer (SLA), which specifies acceptable Quality of Service (QoS) metric levels. From the view of cloud-based software developers, their application-level SLA must be mapped to provided virtual resource-level SLA. Hence, one of the important challenges in clouds today is to improve QoS of computing resources. In this paper, we focus on developing a comprehensive autoscaling solution for clouds based on forecasting resource consumption in advance and validating prediction-based scaling decisions. Our prediction model takes all advantages of fuzzy approach, genetic algorithm and neural network to process historical monitoring time series data. After that the scaling decisions are validated and adapted through evaluating SLA violations. Our solution is tested on real workload data generated from Google data center. The achieved results show significant efficiency and feasibility of our model.

Peer-review under responsibility of the scientific committee of the International Conference on Computational Science

Keywords: Fuzzy time series, neural network, genetic algorithm, adaptive resource management, decision model, SLA, autoscaling, cloud computing

1 Introduction

Thanks to virtualization, clouds can allocate computational resources quickly and dynamically. From the view of software-as-a-service (SaaS) developers, they must ensure the quality of service (QoS) for their end-users. However, the developers are depended on resource QoS provided by cloud infrastructure vendors. This QoS relation among cloud resource providers, software developers and end-users leads to appear Service Layer Agreement (SLA), which is a contract specifying the quality expectation of provided cloud services. Therefore, resource QoS guarantee and SLA violation prevention are crucial points for SaaS developers.

*Corresponding author

Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.121

In fact, most of Infrastructure-as-a-Service (IaaS) clouds offer at least one resource monitoring solution for customers, who can rely on collected data and thresholds to decide amount of resources and scaling moments themselves. The main drawback of the approach is that cloud systems often response quite slowly in comparison with actual requirements from applications, especially with sudden demands. Besides, wasting and lacking resources problems will occur because it is difficult to determine exactly the scaling moments using the threshold technique. In recent years, extensive efforts have been conducted in the area of resource QoS improvement and SLA audit for cloud systems. In order to enhance QoS of resource provision, there are many studies that have dealt with building prediction models [17], [8] for application resource consumption. However, while the studies concern with the forecasting problem in the manner of improving accuracy among models together, they lack methods to evaluate the model effectiveness when making resource increase or decrease decisions based on achieved predictive results. In other words, they do not provide any solution to validate the prediction models in scaling process. In the aspect of SLA, many languages and frameworks [2] and [1] have been developed to keep close control of SLA violations. Unfortunately, these related SLA proposals only focus on auditing QoS based on traditional resource monitoring. In the scenario of applying prediction models to scale resources in cloud systems, there is still a need of having a solution to prevent SLA violations. In those directions, our work described in this paper has the following contributions: (1) Building a novel proactive autoscaling model for clouds includes two main components: prediction and scaling decision; (2) Proposing a novel prediction approach that exploits simultaneously multiple monitoring utilization data such as CPU, memory to forecast the future system usages; (3) Applying fuzzy time series approach in the prediction model to improve forecast effect; (4) Proposing a novel approach to make scaling decisions based on multiple parameters including predictions of multi-resource utilization and SLA violation estimation.

The rest of this paper is organized as follows. In section 2, we discuss some related studies to highlight the differences between our work and existing researches. In section 3, we present the model design of our proactive autoscaling system. Section 4 presents our experiments, gained results and observations with our proposals using real Google cluster data [13] to demonstrate the efficiency and feasibility of the model in practice. Finally, section 5 concludes and figures out some future directions.

2 Related work

Recently, problem of dynamic resource provisioning for clouds has been studied extensively. A lot of efforts deals with trade-off between minimizing resource consumption and meeting SLA. These provisioning techniques could be classified roughly into two categories: reactive and predictive provisioning.

The first category focuses on resource allocation methods [7], which react to immediate demands. In this way, Nguyen et al. [10, 11] propose a three-state model for server management that adjust resources by turning on or off servers according to job arrival and departure. In [4], Dutreilh and et al. apply threshold-based policies for autoscaling actions to adapt resources according to requirements. Thus, resources are allocated or deallocated to applications if performance metrics pass the upper or lower thresholds. Authors of [6] use a set of four thresholds and two duration of multiple performance metrics to trigger resource scaling actions. However, using these techniques, due to delay during adding or removing resources, scaling actions might not achieve the desired effect as compared with actual application requirements. Moreover, this also may lead to SLA violation phenomenon.