Proactive Auto-Scaling Algorithm (PASA) for Cloud Application

Mohammad Sadegh Aslanpour, Department of Computer Engineering, Sirjan Branch, Islamic Azad University, Sirjan, Iran
Seyed Ebrahim Dashti, Department of Computer Engineering, Jahrom Branch, Islamic Azad University, Jahrom, Iran

ABSTRACT

Application providers (APs) leave their application hosting to cloud with the aim of reducing infrastructure purchase and maintenance costs. However, variation in the arrival rate of user application requests on the one hand, and the attractive cloud resource auto-scaling feature on the other hand, has made APs consider further savings in the cost of renting resources. Researchers generally seek to select parameters for scaling decision making, while it seems that analysis of the parameter history is more effective. This paper presents a proactive auto-scaling algorithm (PASA) equipped with a heuristic predictor. The predictor analyzes history with the help of the following techniques: (1) double exponential smoothing - DES, (2) weighted moving average - WMA and (3) Fibonacci numbers. The results of PASA simulation in CloudSim is indicative of its effectiveness in a way that the algorithm can reduce the AP’s cost while maintaining web user satisfaction.

KEYWORDS

Auto-Scaling, Cloud Computing, CloudSim, Proactive, Service Level Agreement

1. INTRODUCTION

With the advent of cloud computing, the Internet business became more appealing. Cloud providers (CP) rent computing infrastructures to customers in a competitive market in the form of pay per use. Offering virtual machines (VMs) in its Elastic Compute Cloud - EC2 service, Amazon has won a large share in this type of business. This environment has created the opportunity for web application providers to rent this VM with an approach to decrease the costs of buying and maintaining computing infrastructure; the provided application can be an online store. However, nowadays, the economic expectations of the AP from the CP is more than merely application hosting; the reason is that the AP is faced with the issue of variation in the arrival rate of end user requests and unawareness of the required resources. Thus, it intends to automatically prevent resource over-provisioning and under-provisioning. Over-provisioning imposes costs on the AP and under-provisioning results in end-user dissatisfaction. Thereby, an effective solution to this challenge is to reduce the cost of renting VMs as long as it is able to consider quality of service (QoS) restrictions. In cloud environments, these QoS limitations are recognized under a service level agreement (SLA) contract which is signed between the service provider and the recipient after negotiations on the QoS. Creating a balance between cost and performance is a fundamental issue in Internet business (Jiang, Hsu, & Wang, 2015). To solve this challenge, thanks to the scalability feature of cloud resources, an algorithm can be provided to automatically carry out resource scale-up or down (Veni & Bhanu, 2016).

In terms of the type of dealing with the decision-making parameter, scaling algorithms are divided into two categories: reactive and proactive. Reactive scaling is generally a rule-based (threshold-
based) method which defines limitations for violating a set of rules and carries out resource scaling after observing violation of the rule. But there are three important concerns regarding this method:

- The on-demand VM in Amazon’s EC2 service requires 5 to 15 minutes for startup (Islam, Keung, Lee, & Liu, 2012; Mao & Humphrey, 2012; Kaur & Chana, 2014). Thus, after the violation of the threshold, the scaling puts the AP on hold while the VM is being run.
- Decision-making after the violation of the rules is associated with SLA violation penalty for the AP (Qu, Calheiros, & Buyya, 2016).
- Some threshold violations could be due to temporary fluctuations and practically, there may be no need for making scaling decisions (Xiao, Song, & Chen, 2013).

However, the other method in determining the scaling algorithm, is the proactive method which seem to be able to cover the violations of the reactive approach by considering the history and predicting the future. This method is able to detect the resources required in the future before crisis conditions and SLA violation. In addition, it can overcome the challenge of delayed VM start-up by predicting the future.

On the other hand, researchers are generally trying to show what scaling parameter can be effective in scaling decision-making; as the authors of this study addressed the issue of selecting resource utilization parameters and response time (RT) in the previous work (Aslanpour & Dashti, 2016). Nevertheless, it seems that knowledge discovery and analysis from the history of each parameter can have an effective role in achieving efficient AS. Today, the analysis of the data created in the web environment is among important (Aye & Thein, 2015). Therefore, the aim of this study is to present a proactive auto-scaling algorithm (PASA) to discover and analyze knowledge from the history of the decision-making parameter. The scientific contribution of this article is as follows:

- Heuristic prediction algorithm to discover and analyze knowledge from the history of the decision-making parameter
- Proactive auto-scaling algorithm (PASA) based on the MAPE loop
- Resource-aware and SLA-aware auto-scaling for cloud applications
- Designing resource auto-scaling algorithm in CloudSim for web application

Related literature is reviewed in the following part. Next, the proposed prediction algorithm is presented. Afterward, proactive auto-scaling algorithms are presented and then, we have the simulation of PASA in CloudSim and the evaluation of its performance. The final part is the conclusion.

2. RELATED WORKS

Resource AS in cloud computing has been discussed from different points of view. Proactive and reactive methods can be the most outstanding feature for categorizing research.

2.1. Reactive Methods

Mohamed et al. (2014) conducted a study on SLA-aware resource scaling with reactive method without studying resource status. They also failed to analyze the criteria related to resource utilization, cost and SLA violation. Garcia, Espert and Garcia (2014) presented an architecture for the management of resources equipped with horizontal scaling algorithm. The proposed method benefits from SLA awareness is in the form of RT. In the evaluation of the performance of their algorithm, they worked on cost without considering SLA violation penalty. In another study, Molto, Caballer and de Alfonso (2016) presented a resource management architecture which performs auto-scaling using a resource-aware method. The analyses this algorithm is provided with for decision-making is only aware of
Performance Evaluation of WMN-GA Simulation System for Different Settings of Genetic Operators Considering Giant Component and Number of Covered Users
[www.igi-global.com/article/performance-evaluation-wmn-simulation-system/67554?camid=4v1a](www.igi-global.com/article/performance-evaluation-wmn-simulation-system/67554?camid=4v1a)

The Development of ICT for Envisioning Cloud Computing and Innovation in South Asia
[www.igi-global.com/chapter/development-ict-envisioning-cloud-computing/64496?camid=4v1a](www.igi-global.com/chapter/development-ict-envisioning-cloud-computing/64496?camid=4v1a)