International Journal of Grid and High Performance Computing
Volume 12 • Issue 1 • January-March 2020

Auto-Scaling Provision Basing on Workload Prediction in the Virtualized Data Center

Danqing Feng, Harbin Institute of Technology, Harbin, China & AirForce Communication NCO Academy, Dalian, China

Zhibo Wu, Harbin Institute of Technology, Harbin, China

Decheng Zuo, Harbin Institute of Technology, Harbin, China

Zhan Zhang, Harbin Institute of Technology, Harbin, China

ABSTRACT

With the development in the Cloud datacenters, the purpose of the efficient resource allocation is to meet the demand of the users instantly with the minimum rent cost. Thus, the elastic resource allocation strategy is usually combined with the prediction technology. This article proposes a novel predict method combination forecast technique, including both exponential smoothing (ES) and auto-regressive and polynomial fitting (PF) model. The aim of combination prediction is to achieve an efficient forecast technique according to the periodic and random feature of the workload and meet the application service level agreement (SLA) with the minimum cost. Moreover, the ES prediction with PSO algorithm gives a fine-grained scaling up and down the resources combining the heuristic algorithm in the future. APWP would solve the periodical or hybrid fluctuation of the workload in the cloud data centers. Finally, experiments improve that the combined prediction model meets the SLA with the better precision accuracy with the minimum renting cost.

KEYWORDS

ES, PF, Prediction, Provisioning, Scaling, SLA

INTRODUCTION

Recently, the two main features of the cloud computing are both the elasticity (Herbst, Kounev, & Reussner, 2013) and virtualization (Yang et al., 2009) in the cloud data centers. That is to say, it is important to meet the fluctuating demand (Weingärtner, Bräscher, & Westphall, 2015) of the users. However, static lower or upper threshold would lead the poorer utilization and the more energy consumption. If the resources are under-provisioning, it would not meet the demand of the users, and cause the punishment by the SLA. If the resources are over-provisioning, it is another emergent problem to solve for the energy consumption (Chihoub et al., 2015) and heat loss.

To make the further fine-grained scaling in the resource provisioning, the auto-scaling provisioning (Roy et al., 2011) used to be combined with the predict technique. The purpose of it is to avoid excessive or inadequate resources to be allocated. The under-provisioning resources would not meet the demand of the users, which causes SLA violations (Musa et al., 2016). On the contrary, the over-provisioning resources would cause the resources waste during the deploying process in the

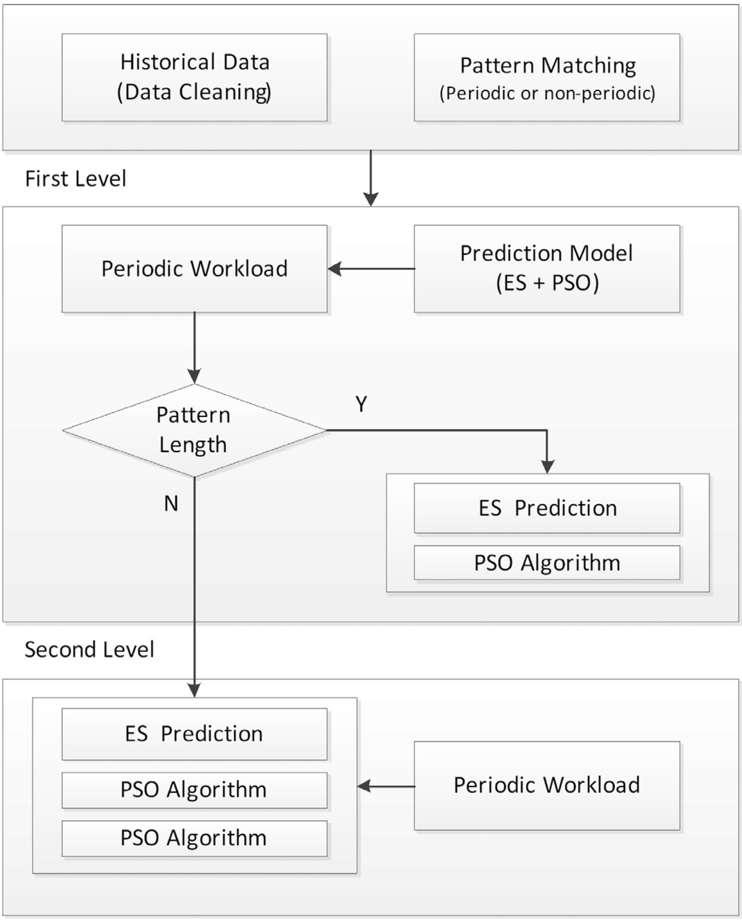
DOI: 10.4018/IJGHPC.2020010104

cloud data centers. Thus, the two challenges in the allocating resources of the cloud computing is to solve the following problem: (1) deciding the efficient elastic provisioning (Dustdar et al., 2011) in the data centers, (2) improving the prediction accuracy and reducing the time complexity, (3) minimizing the renting cost and service level objective (SLO) violations (Manvi & Shyam, 2014).

Therefore, in this paper, we propose a novel combined prediction technique as a kind of efficient resource provisioning strategy in the cloud data centers, as it is shown in Figure 1. The combined prediction algorithm is mainly composed of two parts. The basic algorithm is the ES model with PSO algorithm due to its flexible calculation basing on a few samples. Then, the reactive prediction model is PF model just for reducing the under-provisioning statements. The aim of the combined prediction model is to minimize the renting cost and the SLO violations (Hwang et al., 2013). We analyze and determine the features of the demand which is composed of repeated patterns or unrepeated ones. The purpose of the proposed model is to improve the accuracy and minimize the overheads. According to the analysis, we present the two-level forecasting technique:

- In the first level, we choose the ES model to predict the varying demand with the workload. Single exponential smoothing curve is fit to forecast closer to actual observations by setting different weights. Then, the weight is based on the calculation of the prediction accuracy. If the time series have the rising or falling trend, the prediction deviation would be less prediction accuracy, then it

Figure 1. Two-Level prediction process


Historical Data
(Data Cleaning)
Pattern Matching
(Periodic or non-periodic)
First Level
Periodic Workload
Prediction Model
(ES + PSO)
Pattern
Length
Y
N
ES Prediction
PSO Algorithm
Second Level
ES Prediction
PSO Algorithm
PSO Algorithm
Periodic Workload