Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 130 (2018) 683–688

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Ambient Systems, Networks and Technologies, ANT-2018 and the 8th International Conference on Sustainable Energy Information Technology, SEIT 2018, 8-11 May, 2018, Porto, Portugal

Virtual Machine Classification-based Approach to Enhanced Workload Balancing for Cloud Computing Applications

Mousa Elrotub[a], Abdelouahed Gherbi[b]

Department of Software and IT Engineering, École de Technologie Supérieure(ETS), Montreal, Canada

Abstract

Despite the many researches that have been conducted in the field of Cloud computing, it is still facing some issues and challenges, such as load balancing which still needs more optimizing methodologies and models to improve performance and achieve high user satisfaction. In this paper, Machine Learning Technique, which is classification, is used to make groups of VMs based on their CPU and RAM utilization, as well as to classify user jobs/tasks into different groups based on their sizes and based on information from log files. The approach arranges virtual machines in groups, and several tasks share the same VM resources. The goal of our proposal is to allow more dynamic resources and to improve the QoS requirements by maximizing the usage of the resources and user satisfaction, such as increasing resource utilization and reducing the number of job rejections.

Keywords: Cloud Computing; QoS; Load Balancing; Classification; Log files, Machine learning

1. Introduction

Cloud computing basically provides shared computing and storage resources. Also, one of the requirements of this model is scheduling and allocating the current jobs/tasks to be executed with a high level of Quality of Services (QoS) which users need and maximizing cloud resource usage. The increased transitions to a SaaS model with the variations in the pattern of access to the application based on specific periods are an important issue. There are often many users who might be trying to use the service at the same time. Moreover, some of them are very likely to access the concurrent servces[3].

The log files are an essential source of information. More specifically, it can be used to retrieve valuable information about the users who access the web from different backgrounds. Analyzing and modeling web navigation behavior is helpful in understanding online users' demands. Such insight can be used to enhance the

Corresponding author. Tel.: +1-514-992-7589 ; fax: +1-514-840-5514. E-mail address: mousa.elrotub.1@ens.etsmtl.ca

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2018.04.120

efficiency and performance of Job Scheduling tasks.

In this paper, we propose and present a model to address tasks workload concern and how to support automatic handling of each VM and to find the appropriate VM for each requested task. The main contributions of this paper include the followings:

- To propose the design of a model which uses the information from log files to classify user tasks based on their sizes and then calculating the task's resources requirements. This allows for an efficient allocation of tasks to the VMs in Data Centers in order to maximize resource usage.
- To present a new approach for identifying each VM capacity by calculating its CPU and RAM utilization as percentages (%) for maximizing the number of tasks that can be placed in each VM.
- To discuss how to leverage one of machine learning techniques which is classification to make groups of VMs based on their resources' usage to facilitate a virtual machine (VM) placement.

The remainder part of the paper is organized as follows: In Section 2, we describe the problem definition and limitations. We present in Section 3 the previous studies and related work. In Section 4, we outline our methodology which is followed by the solution. Section 5 explains the proposed system and its architecture using classification technique. Section 6 presents a further discussion of our approach. Finally, we conclude the paper and present our future work in section 7.

2. Problem Specification

Balancing workload of incoming user requests and allocating the corresponding tasks to appropriate virtual machines in a virtual environment is a challenging issue. This is further exacerbated considering that the efficient workload management becomes more complicated with the increasing number of requests with an unpredictable arrival pattern. In order to overcome this issue, there is a need for efficient and high-quality methodologies and models to support improving the cloud-based application's performance by optimizing the system resources' usage and therefore raising the end-user satisfaction. Examples of performance attributes include the reduced response time and the number of job rejections.

In this paper, we limit the scope of our focus mainly on the workload relevant for the allocation of the tasks in each DC/partition. We also identify and target some limitations with the established load balancing techniques. These limitations are classified as follows:

- Numerous works in the literature[1, 4, 6, 7, 9] focus either on CPU utilization ratio or RAM utilization ratio but not both. Also, they usually check if the VMs are busy or not and do not measure them in current usage percentages, ranging from 0% to 100% for determining how much they are used.
- Log files are a significant repository of information, and there are different types of users, which are not considered;
- The priority of the jobs is not a factor that is considered by most researchers.
- Most studies use the information from log files to predict resource usage and not for predicting incoming requests.

3. Related Work

Several approaches in various studies have been proposed using clustering technique and log files but still some limitations. In[1], the authors propose architecture by using K-Means method based on clustering VMs to place them in each Data Center (DC). The attribute that is considered in the algorithm for clustering is just the virtual machine's RAM size. They use a mathematical model for explaining the concept of the proposed system. In[2], the authors' proposed architecture uses clustering techniques for cloud resource allocation. They use Google cluster traces to make groups of tasks and map them to the virtual machines, taking into consideration the actual resource utilization of each cluster, and do not focus on the amount of resources that are requested by the users. The mapping process is based on the task usage patterns, which are obtained from the historical data. In[5], the authors propose a