Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 717–724

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

Elastic and flexible deadline constraint load Balancing algorithm for Cloud Computing

Mohit Kumar[a*], Kalka Dubey[b], S.C.Sharma[a,b]

[a]*Research Scholar, IIT Roorkee, india*
[a,b]*Professor, IIT Roorkee, India*

Abstract

There are many load balancing algorithm has been proposed for cloud computing in last decade but none of algorithm provides the elasticity with load balancing. We proposed a cloud architecture that is capable of handling the maximum user request before meet to deadline and provides an elasticity mechanism with the help of threshold based trigger strategy. Computational results (Table 1 & Figs. 2-5) shows that develop algorithm decrease the makespan time and enhance task acceptance ratio more than 10% compare to min-min algorithm, 30% compare to First come first serve (FCFS) and shortest job first (SJF) in all condition.

Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications.

Keywords: scalability, makespan time, virtual machine, elasticity, task scheduling;

1. Introduction

Cloud computing is an internet based computing technology in the field of computer science that provides the services (either in the form of software or hardware) to the users on the basis of pay per use. It provides the services like as software as a service, storage as a service, infrastructure as a service etc.[1] and three deployment models of cloud are public, private and hybrid. Public cloud services are available for general public over the internet. Amazon elastic compute cloud, Google appEngine, Window azure service platform etc are examples of public cloud those offered the services either pay per use basis or free. Private cloud is used for personal use or provides the service to

* Corresponding author: Tel. +91-9759950380
E-mail address: mohit05cs33@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.092

single organization. Eucalyptus, OpenNebula, Openstack etc are example of private cloud those offered the similar advantages to public cloud. A hybrid cloud is combination of two or more than two public and private cloud which are bounded by service level agreement (SLA). User can send the request at any time from any geographical location for services, SLA selects the best resource within user defined deadline and budget. Elastic resource provisioning with quality of service (QoS) parameter (deadline, high availability, priority etc.) is one of the most challenging problem in the field of cloud computing. Therefore cloud service provider needs an efficient load balancing algorithm that reduces the makespan time as well as task rejection ratio within user defined deadline.

In the last few years, maximum company are trying to achieve the scalability in terms of platform, application and infrastructure level. Scalability is an important feature in cloud computing and can be divided into two part one is scale up other is scale out [2]. Scale up is also called vertical scalability and scale out is also called horizontal scalability. In this paper, we are using the horizontal scaling approach based upon the workload prediction with the help of user defined threshold at the time of service level agreement (SLA).

The reminder of the article is organized as follows: Section 2 describes the related work which is related to our research work such as existing scheduling & load balancing technique with virtual machine provisioning and deprovisioning, Section 3 we will discuss the proposed architecture and its components, section 4 problem formulation and proposed scheduling algorithm with elasticity, further section 5 is for analyze and comparison of experimental results and Section 6 conclusion.

2. Related Work

There are lot of algorithm have been proposed for load balancing and scalability of cloud resource in last decade. E.Coninck et al. proposed a dynamic auto scaling algorithm that reduces the execution time and makespan time of upcoming requests (application/task) considering the deadline as a constraint using the Openstack and cloudsim as a simulator [3]. M. Kumar and S.C Sharma [4] proposed an algorithm that not only reduce the makespan time of tasks but also increase the utilization ratio of the task considering the priority of tasks as quality of services parameter. F.Juarez proposed a dynamic energy aware scheduling algorithm that reduces the makespan time and energy using the private cloud as a tool for implementation [5]. In this paper, find out the makespan and energy consumption for results but integration of both parameters doesn't give optimal results i.e. one parameter at a time because trade-off occur between time and energy. Ye Feng et al., proposed a dynamic load balancing algorithm that reduce the task completion time and load balancing degree [6]. S. Abrishami, M. Naghibzadeh develop an algorithm for SaaS and IaaS that reduce the parameter cost and execution time where deadline as a constraint and java based simulator is used to implement the algorithm[7][8]. Proposed algorithm is implemented on Java based simulator that does not give the guarantee of cloud environment. The most important problem in the real environment is the inaccuracy of the estimated execution and transmission times. R.Naha, M.Othman and T.Somasundaram proposed broker based architecture for task scheduling and elasticity in cloud environment using cloud analyst and eucalyptus as a tool for implementation [9][10].

Li Xiaofang et al. [11] proposed an improved max-min algorithm for elastic cloud that monitor the load at virtual machine continuously. Proposed algorithm allocate the task to running virtual machine in such a way that it can improve the response time of tasks and resource utilization ratio. Coutinho et al. Analyze the elasticity behaviour in cloud computing [12] using the parameter response time and utilization of cpu. Further author's define the over provisioning and under provisioning condition in elasticity. Al-Dhuraibi, Yahya, et al. Review all the existing classical and recent elasticity solutions [13]. Author's compared the horizontal vs. vertical scalability pros and cons. Further they discuss the classification of elasticity mechanism, performance evolutions tools and elasticity solution in details. Galante et al. [14] present a comprehensive study of elasticity including the classification mechanism based upon the commercial and academic solutions in cloud environment. Some challenges and open issues which are associated with the use of elasticity concept are also discussed by the author's in this paper. Hu, Yazhou, et al. [15] discussed the elasticity concept and proposed a linear regression model to predict the upcoming workload in