CHAPTER I

INTRODUCTION

1.1 Background

Stryer (2010, 4) define cloud computing as a computational model which allows computation to be run on a remote data center, thereby reducing resource usage on the end user device. The architecture of cloud computing has its unique characteristics in which utilizes concepts of virtualization, processing power, storage, connectivity, and sharing to provide resources elastically and with an on-demand basis (Moghaddam 2015, 1). The need for proper workload balancing arise as there are significant increase on data traffic volume and user demand for computational resources during peak time, which in turn causes resources to be fully utilized and leaving no room for additional requests to be handled. This problem is commonly referred as server overloading (Afzal & Kavitha 2018, 374). A study conducted by Iyer et al. (2000, 237) observed on how overloading negatively impacts the performance of a server, most noticeably as an increase on response time to handle incoming client requests which corresponds to longer request queue lengths.

In cloud computing, principles of elasticity aims to overcome overloading issues by scaling system resources to increase or decrease capacity according to traffic demand or workload. Herbst et al. (2013, 2) defined elasticity as:

> the degree to which a system is able to adapt to workload changes by provisioning and deprovisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as closely as possible.

A solution to achieve maximized elasticity in cloud computing is to implement a load balancer system configured with an algorithm that has the best performance to balance incoming workloads, so that there are no overloaded nor underloaded nodes. In practice, implementing a load balancer system makes optimal use of available resources, thus subsequently lower resource consumptions, improve scalability, and ultimately reducing response time (Kansal & Chana 2012, 239).

An effort has been done in previous research to compare cloud computing load balancer algorithms, such as researches conducted by Shakir & Razzaque (2017, 509), Rajeshkannan & Aramudhan (2016, 1) and Prasetijo et al. (2016, 393). However, the results of such experiments are interpreted solely by descriptive observation and may not produce the most accurate conclusions. With data analytics, it is possible to fulfill this research gap by making use of a predictive data analytic approach to determine a load balancer algorithm with the best performance based on various workload characteristics.

This research aims to compare the performances of Round Robin, Least Connection, and Weighted Least Connections algorithms. The three algorithms are chosen to represent static, dynamic, and weighted dynamic load balancing algorithm characteristics respectively. To simulate real server traffic conditions, various levels of workloads will be applied onto the load balancer using Apache JMeter. The performance results of each algorithm will then be captured and analysed using methods of data analytics to determine which algorithm has the best performance under various simulated workloads.

1.2 Problem Formulation

The formulation of the problems of this research is as follows:

1) How does a load balancer perform under various workloads within a virtualized environment?
2) What is the algorithm that has the best performance that could be implemented on a load balancer within a virtualized environment under various workload characteristics?

1.3 Scope Limitations

There are limitations and restrictions applied upon conducting this research, which as follows:

1) Load balancer is run within a virtualized environment on Amazon Web Services
2) Workload is simulated using Apache JMeter
3) Algorithms used are Round Robin, Least Connection, and Weighted Least Connection
4) Captured data is analysed using data analytics and machine learning

1.4 Purpose

This research aims to observe performance and behaviour of load balancer algorithms under various workloads within a virtualized environment. The observation is conducted through comparing the performances of Round Robin, Least Connection, and Weighted Least Connections algorithm to determine a load balancer algorithm with the best performance under various workload characteristics. The performance results of each algorithm will be captured and analysed using data analytics. The expected result of this research is to determine a load balancer algorithm which has the best performance under various workloads, in which can be implemented on a real cloud server to maximize computing elasticity and provide better user satisfaction.

1.5 Methodology

1) Conducting literature study on the implementation of Round Robin, Least Connection, and Weighted Least Connections algorithms on a load balancer
2) Configuring virtualized environment

   a. Running a virtualized environment on Amazon Web Services

   b. Configuring load balancer using HAProxy

   c. Configuring workload simulation using Apache JMeter
3) Implementing algorithms to load balancer
4) Running experiments to capture data
5) Analysing results to draw conclusions