Proposing A Load Balancing Algorithm For The Optimization Of Cloud Computing Applications

Dalia Abdulkareem Shafiq
School of Computing & IT (SoCIT), Taylor's University
Subang Jaya, Malaysia
daliakareem7@gmail.com

NZ Jhanjhi
School of Computing & IT (SoCIT), Taylor's University
Subang Jaya, Malaysia
noorzaman.jhanjhi@taylors.edu.my

Azween Abdullah
School of Computing & IT (SoCIT), Taylor's University
Subang Jaya, Malaysia
Azween.Abdullah@taylors.edu.my

***Abstract*—Cloud Computing (CC) is a fast growing services that make use of pay per use model. The technology provides various services in terms of storage, deployment, web services etc. however the expand of these services and the tremendous increase of user demand has resulted in many challenges to keep up the performance in line with QoS measurement and SLA document provided by cloud providers to enterprises. This expand resulted in challenges such as load balancing. Besides that, user's requirements became hard to fulfil in terms of response time and deadline regarding task scheduling. To address these challenges, this research proposes an optimized algorithm with the use of Machine Learning Classification technique based on deadline constraints. The main objective of the proposed algorithm is to enhance the efficiency, optimize the server resources by considering the priority of different users' tasks and avoid server breakdown. Our proposed algorithm will address the mentioned issues and current research gap based on the recent literature.**

Keywords—Cloud Computing; Virtualization; Task Scheduling; Load balancing; Machine Learning; Classification; Optimization.

I. INTRODUCTION

Cloud Computing (CC) is an emerged technology that provides services for storing, accessing files and data over the cloud rather than locally on computers. A term proposed by Prof. Ramnath Chellapa in 1997 [1] and it's defined [2] as a dynamic system architecture designed to provide inexpensive various services over the internet to clients. The technology aims to enhance business worldwide with its scalable environment and reduced costs in terms of hardware. CC has three delivery models namely: Platform as Service (PaaS), Software as Service (SaaS) and Infrastructure as service (IaaS). In SaaS the services are accessed through web browsers such as google docs, Gmail and other services. Cloud providers also support the building of client services by providing platforms and programming languages this is PaaS. Lastly example of IaaS is data centres which is storage of data, where clients have less control over the underlying infrastructure and more control in terms of operating system (OS). The pay per use services are provided by big IT companies such as Google and Microsoft. Among the three service models, it is found that SaaS is mostly used by organizations [3], this is due to its simplicity as it does not require installation and services can be easily accessed through web browsers.

Cloud services have seen a tremendous growth over the years. Based on statistical fact from 451 Research, currently in 2019, 60% of the companies' workload are on the cloud whereas in 2018, the workload was 45% [4]. This fact proves that there's a remarkable growth in the utilization of cloud services by companies. Such expand in services often raise challenges to cloud providers to keep up the quality of services provided to their clients. According to [5] performance is one of the top three challenges of CC. Challenges such as load balancing could degrade the performance of CC applications and this lower the user satisfaction rate.

Virtualization is an essential feature of CC applications when compared to other technologies such as grid computing and utility computing [1]. It converts physical components such as OS, servers, storage devices etc. into a virtual containers known as Virtual Machines (VMs). In another words, virtualization creates an abstract layer between software and hardware [6]. This is done with the help of hypervisors which is sometimes referred to as Virtual Machine Monitor (VMM) to allow multiple VMs to run on a single hardware layer. As can be seen in the fig. 1 below, there are two types of VMM, type 1 directly interacts with the hardware layer whereas for type 2 a host OS is required to provide services for support, storage etc. Each VM contains two layers: Guest OS (Window, Linux and Mac) and Application layer. With this feature, CC can provide on-demand services to clients that are more scalable. Since virtualization technology plays a vital role in CC, it can greatly affect the performance of applications if there's inefficiency in virtual machine migration and task allocation.

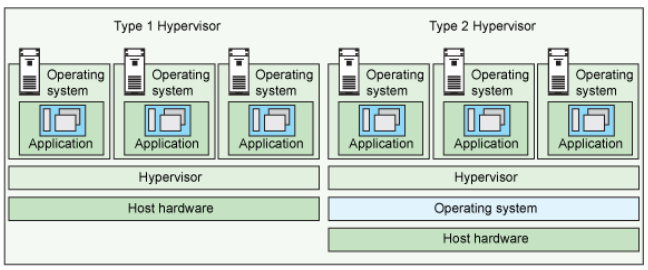

Fig. 1. Types of Hypervisors [7].

978-1-7281-4956-1/19/$31.00 ©2019 IEEE