International Journal of Software Innovation
Volume 8 • Issue 1 • January-March 2020

Dynamic Management of Resources in Cloud Computing

Pradeep Kumar Tiwari, Manipal University Jaipur, Jaipur, India
https://orcid.org/0000-0003-0387-9236

Sandeep Joshi, Manipal University Jaipur, Jaipur, India

ABSTRACT

It has already been proven that VMs are over-utilized in the initial stages and are underutilized in the later stages. Due to the random utilization of the CPU, resources are sometimes heavily loaded whereas other resources are idle. Load imbalance causes service level agreement (SLA) violations resulting in poor quality of service (QoS) aided by the imperfect management of resources. An effective load balancing mechanism helps to achieve balanced utilization, which maximizes the throughput, availability, and reliability and reduces the response and migration time. The proposed algorithm can effectively minimize the response and the migration time and maximize reliability, and throughput. This research also helps to understand the load balancing policies and analysis of other research works.

KEYWORDS

Migration, Reliability, Scalability, Throughput

1. INTRODUCTION

Cloud Computing is an internet-based computing service in which user can access the application and computing resources via internet. The management of demanded resources is known as load balancing mechanism in which workload is distributed among the virtual machines (VMs). Load Balancing mechanism is a key component of the hypervisor, which dynamically or static manage the load imbalance in distributed manner on the available VMs. CPU, memory and network components are virtualized to maximize the utilization of resources. (Joseph, Chandrasekaran & Cyriac, 2015).

1.2. Load Balancing Mechanism

The Cloud system is rendered ineffective by the load imbalance, which is also caused due to poor availability of resources, reliability, scalability, and throughput. In order to enhance the reaction time of the employment and to compel the asset utilization, the total work load is reassigned to individual hubs on the framework. Such a process is termed as load balancing, which nullifies the situations in which the hubs are either under-stacked or over-stacked. Hence, load adjusting is generally a system that encourages systems and assets by giving a maximum throughput at the least reaction time by partitioning the movement between servers. Load adjusting calculations can be fundamentally

DOI: 10.4018/IJSI.2020010104

classified into static, dynamic or symmetrically managed load balancing (Singh, Juneja & Malhotra, 2015; Buyya, Ranjan & Calheiros, 2010).

Availability of physical and logical components in physical computing machine is known as resources. Cloud computing uses physical (i.e. CPU, memory, secondary Storage, Work Station) and logical resources (i.e. operating system, energy, network throughput, load balancing mechanism) (Xiao, 2015).

Load balancing mechanism not only manages the load distribution among the available VMs, but also controls the load imbalance with a fault tolerance. The response time is minimized and the throughput is maximized by effective load balancing. Figure 1 shows the load balancing policies, which play a vital role in distributing the fair load among the VMs.

1.2.1. Transfer Policy

Transfer policy is based on the CPU's threshold state. It can be gauged from a high threshold that jobs need migration, since none of them are being executed by the CPU. On the contrary, it can be gauged from a low threshold that the current CPU is capable of executing more loads and is anticipating the load from a high loaded VM.

1.2.2. Selection Policy

Selection policy selects high- and low-loaded VMs. The selection policy may be either static or dynamic, which will select the best fit low-load VM to migrate the jobs of high load VMs.

1.2.3. Location Policy

Location policy identifies the location of a high load VM to migrate the jobs to a low load VM. This mechanism is based on the timeout of CPU. Less timeout indicates that CPU is free to take more jobs from long timeout CPU.

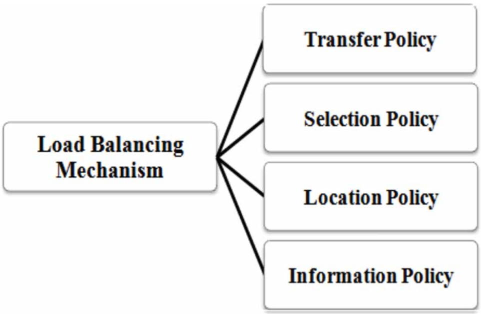
1.2.4. Information Policy

Information policy has the resource information of the available VMs. It has the data centers (DCs) and the information on available VMs, which helps in mapping the VMs' resources to DCs. The manager separates the high and low loaded VMs to find the sender and receiver VMs. The information policy refreshes the dashboard information after every completed migration of jobs (Sammy, Shengbing & Wilson, 2012; Lau, Lu & Leung, 2006; Lu & Lau, 1995).

These load balancing policies are interrelated to each other for managing the user base (UB) request on the available VMs and transferring the job from a high-loaded VM to a low-loaded VM.

Load imbalance is caused because of poor throughput, scalability, reliability, and availability of resources. Load imbalance maximizes the service level agreement (SLA) violations and migration

Figure 1. Load balancing policies


Load Balancing Mechanism
Transfer Policy
Selection Policy
Location Policy
Information Policy