Chapter 16

Load Balancing in Cloud Computing: Challenges and Management Techniques

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

Geeta Rani
Manipal University Jaipur, India

Tarun Jain
Manipal University Jaipur, India

Ankit Mundra
Manipal University Jaipur, India

Rohit Kumar Gupta
Manipal University Jaipur, India

ABSTRACT

Cloud computing is an effective alternative information technology paradigm with its on-demand resource provisioning and high reliability. This technology has the potential to offer virtualized, distributed, and elastic resources as utilities to users. Cloud computing offers numerous types of computing and storage means by connecting to a vast pool of systems. However, because of its large data handling property, the major issue the technology facing is the load balancing problem. Load balancing is the maximum resource utilization with effective management of load imbalance. This chapter shares information about logical and physical resources, load balancing metrics, challenges and techniques, and also gives some suggestions that could be helpful for future studies.

DOI: 10.4018/978-1-7998-1021-6.ch016
INTRODUCTION

An effective Load balancing mechanism enhances the fair workload distribution among the VMs. Load balancing use the mechanism of hyper threading to use a single processor as multiple processors. Load balancing intends to minimize the resource and maximize the resource utilization. The core concept of effective load balancing technique maximizes the throughput and minimizes the response time with fault tolerance (Rathore and Chana, 2014).

The load balancing mechanism is a key mechanism to manage the User Bases resources request from Data Centers (Zhang, Cheng, and Boutaba, 2010).

Load balancing provide effective management of resources by resource allocation policy using the task scheduling in distributed environment. The load management mechanism should ensure the:

- Resource availability on time to reduce Service Level Agreement (SLA) violation
- Effective resource utilization during the high or low load.
- Cost effective by using effective management of resources.
- Increasing the Quality of Service (QoS) with robust fault tolerance mechanism.

That is, load balancing help to continue the service by implementing fail over in the cases where the failure of one or more component occurs. Maximization of the throughput, minimization of the response time and avoidance of the overload are the other major advantages of the load balancing. Above all, by keeping resource consumption at minimum, the load balancing techniques help to reduce costs and create the enterprises greener.

All these features make load balancing atop a prioritized subject among computer science researchers and numerous different load balancing approaches were proposed by multifarious researchers. The present chapter is conducting an in depth review of the studies regarding the existing load balancing techniques in Cloud networking and attempt to find shortcomings exist in those proposals as a mean to come up with a novel proposal which can overcome these shortcomings. The review also targets the studies which deal with the factors such as, parameters for identifying hotspots, an algorithm that can evaluate how balanced a system is, a prediction algorithm for estimating the workload after a migration has occurred and an algorithm to determine how costly a migration will be.

The operating system and the applications of a VM function autonomously without any mutual interference. VM is migrated without downtime and VM failure does not affect the distribution of resources among VMs. The service level agreement (SLA) and the Quality of Service (QoS) must be managed by the Load Balance (LB) policy. The main causes of SLA violation are scattered data among heterogeneous servers, hot spot, load imbalance, and weak resource management (Rathore et al., 2014). The occurrence of the load imbalance is when the demands in the heterogeneous environments are frequently changing. Load imbalance can be managed through LB between high load and low load machines. The management of LB is difficult on high resource demands that change frequently. The factors that help in the management of LB are information policies, location to migrate VM, selection of VM, and transfer of load (Zhang et al., 2010).

In order to attain High Performance Computing (HPC) and utilization of computing resources, the elementary concept of the distributed system is utilized to perform cluster, grid and cloud computing. The distributed application determines the distributed paradigms. Virtual Machines (VMs) can comprise of separately functioning operating systems and applications. Virtualization is the core concept of resource
Related Content

**Mining for Mutually Exclusive Items in Transaction Databases**
[www.igi-global.com/article/mining-mutually-exclusive-items-transaction/1789?camid=4v1a](www.igi-global.com/article/mining-mutually-exclusive-items-transaction/1789?camid=4v1a)

**Retailer Case Study**
[www.igi-global.com/chapter/retailer-case-study/27343?camid=4v1a](www.igi-global.com/chapter/retailer-case-study/27343?camid=4v1a)

**TODE: An Ontology-Based Model for the Dynamic Population of Web Directories**
[www.igi-global.com/chapter/tode-ontology-based-model-dynamic/7569?camid=4v1a](www.igi-global.com/chapter/tode-ontology-based-model-dynamic/7569?camid=4v1a)

**An Immune Systems Approach for Classifying Mobile Phone Usage**
[www.igi-global.com/article/immune-systems-approach-classifying-mobile/1784?camid=4v1a](www.igi-global.com/article/immune-systems-approach-classifying-mobile/1784?camid=4v1a)