Computers and Electrical Engineering 57 (2017) 199–208

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Effective resource utilization in cloud environment through a dynamic well-organized load balancing algorithm for virtual machines ☆

M. Vanitha [a,*], P. Marikkannu [b]

[a] Department of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India
[b] Department of Information Technology, Anna University Regional Centre, Coimbatore, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 30 March 2016
Revised 1 November 2016
Accepted 1 November 2016
Available online 30 December 2016

Keywords:
Service level agreement (SLA)
Dynamic load balancing
Cloud computing
Genetic algorithm (GA)

ABSTRACT

Cloud computing is used in almost all domains today. Through the use of cloud-based applications, it has become easier for an internet user to make use of the services and resources that are widely available. The cloud service provider undertakes to deliver all the subscribers' requirements as per the service level agreement (SLA). These resources must be well-protected since they are used by many subscribers. There is a constant high level of demand for these resources and services, and it is therefore necessary to balance the loads on the various servers; this is done in order to avoid congestion in the network and to reduce consumption of the capital or resources. This load balancing uses algorithms such as genetic algorithms (GA) in order to distribute the load on the virtual machines (VM) that are available in the network. A novel load balancing method is proposed, involving a well-organized use of resources, which is known as the dynamic well-organized load balancing (DWOLB) algorithm. This is a powerful algorithm for reducing the energy that is consumed in cloud computing.

1. Introduction

The use of cloud computing is increasing on a daily basis, and the loads experienced by cloud servers are also increasing significantly. The demand on resources has therefore risen greatly, since there are large numbers of users of these applications; the popularity of cloud applications means that there are always a large number of customers. This leads to maximization of productivity which in turn leads to a reduction in the response time in the cloud. Starvation of resources should be minimized, in order to maintain an effective service, since this can lead to possible overloads. It is therefore necessary to reduce the load on the server so that all users are provided with equal performance. Dynamic well-organized load balancing (DWOLB) is a new technique for improving performance through a reduction in the use of resources and for ensuring that a large proportion of users can access the cloud computing service.

Dynamic well-organized load balancing makes use of a genetic algorithm in its operation; parameters such as the CPU and the bandwidth are also used to calculate the load that is put on the users, allowing the cost for each of the available virtual machines (VM) to be changed. The load on the central processing unit is lower when less work is carried out by the

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. S. Smys.

* Corresponding author.
E-mail address: nagavanithaa@gmail.com (M. Vanitha).

http://dx.doi.org/10.1016/j.compeleceng.2016.11.001
0045-7906/© 2016 Elsevier Ltd. All rights reserved.

servers. For instance, when a Twitter server is not in use, it consumes about 70 W of energy. For conditions of low server consumption, the energy consumed is about 140 W, and when the server is fully loaded, the energy consumed is about 160 W. From this, it can clearly be seen that there is little difference in the energy consumed. Hence, it is more effective to ensure that the server is always loaded with work, than allowing servers to idle or to take lower loads. The key observation is that the server must always be in the 'on' state and should be fully loaded.

The remainder of this article is structured as follows. Section 2 discusses the current research in this area, while Section 3 describes the proposed system in more detail. Section 4 examines the concept of load balancing in the cloud. Section 5 introduces the migration of virtual machines in real-world conditions. Section 6 presents an analysis of the dynamic well-organized load balancing algorithm. Section 7 presents the experimental results and a discussion of the implementation of the load balancing algorithm, and Section 8 concludes this paper.

2. Related work

Since there is an increase in the use of cloud computing every day, the load on servers has increased to a great extent. The use of load balancing has therefore been the focus of much research, due to the active and adaptable service that is offered by cloud computing. The main aim of researchers has been to enable an equal load balance while maintaining low power consumption, and analysis has been focused either on load balancing or on the usage of energy by the servers in the virtual machine. One of the most widely used types of algorithms is the genetic algorithm, which is used to identify the server present in the virtual machine. An improved adaptive genetic algorithm was used by Xin LU [1] for the discovery of appropriate solutions for all the applications taking place in real time. Jianhau [2] examined CPU utilization with a genetic algorithm to optimize the resources of the virtual machine. The workload is planned in terms of the processing and memory required for execution, which in turn reduces processing time [3]. In addition, parallel processing can be supported by scheduling tasks using a genetic algorithm, as shown by Jianhua [4]. The genetic function is used for each individual task carried out by the server in work by Tianhai Zhao and Sung [5,6]. The parameters used by the author for calculating performance include average machine utilization, throughput, processing cost and the time required for simulation. The genetic algorithm has been modified as necessary to give rise to a novel algorithm.

Power consumption has also been an important research topic. A novel technique for reducing the cost of the total energy consumed is reported by Lee, Wang [7,8]. A controlled theoretical solution gives rise to a method for active volume provisioning, thereby reducing the cost of the total energy consumed. The simulations are based on real-time hints and suggestions obtained from websites such as Google. An automated system was also proposed for an energy-aware server by Brian [9]; the primary aim of this was to reduce the cost of both the server and the energy needed to run the application on the servers [10]. This system can be valuable in determining the appropriate servers for an application. In addition, server power can be managed at the ensemble layer. An algorithm for distribution of workload has also been provided by Abbasi [11], in which the active servers are selected by users, and a specific threshold is used. This threshold is set so that it does not rise beyond a certain value in the server, in order to ensure correct operation [12,13].

3. Proposed system

The proposed system for dynamic well-organized load balancing is implemented using a VMware workstation. Four vSphere servers are considered, where each server contains four virtual machines loaded with the Windows operating system. Of these four servers, one acts as the management server, and a second as the storage server. Through the use of more active number of VMware servers, movement is enabled from one physical server to another physical level. The Microsoft Hyper-V is 5.4 times slower than the VMware VMotion. The proposed system architecture includes virtual server management, virtual equipment, VMware server, operating system and server. Fig. 1 shows the architecture of the proposed system.

4. Load balancing

Load balancing is essentially a procedure which shares work between two or more entities, in order to obtain the best utilization of resources. This in turn gives the maximum throughput; that is, the amount of data that are transmitted successfully is optimized [14]. There are various internet-based services for load balancing, such as client-side random load balancing, round robin DNS and server-side load balancing.

Similarly to its use within telecommunication, load balancing can be used for bridging the shortest path and for routing purposes. The primary aim of the DWOLB algorithm is to provide optimum resource consumption and throughput, and to reduce response time [15,16]. Hence it becomes necessary to update the load on each server periodically. When used as an algorithm for load balancing, information from the system as a whole is considered and polled before taking a decision regarding the workload on a single server. The DWOLB algorithm is based on a genetic algorithm. Through the use of the genetic algorithm, the decision is made and the final decision is then sent to the other servers. This is done to change the position of the virtual machine with respect to the decisions which have been taken [17,18].