International Journal of Multidisciplinary Research and Modern Education (IJMRME)
ISSN (Online): 2454 - 6119
(www.rdmodernresearch.com) Volume II, Issue II, 2016

A SURVEY ON LOAD BALANCING IN CLOUD COMPUTING USING ARTIFICIAL INTELLIGENCE TECHNIQUES

Amandeep Kaur* & Pooja Nagpal**

* Research Scholar, Department of CSE, I.K Gujral Punjab Technical University, Rayat Institute of Engineering and Information Technology, Punjab

** Faculty of Computer Science and Information Technology, I.K Gujral Punjab Technical University, Rayat Institute of Engineering and Information Technology, Punjab

Abstract:

Since its inception, the cloud computing paradigm has gained the widespread popularity in the industry and academia. The economical, scalable, expedient, ubiquitous, and on-demand access to shared resources are some of the characteristics of the cloud that have resulted in shifting the business processes to the cloud. The cloud computing attracts the attention of research community due to its potential to provide tremendous benefits to the industry and the community. But with the increasing demand of the cloud computing, there are some challenges also. The main cloud computing challenges are Data Management and Resource Allocation, Security and Privacy, Load Balancing, Scalability and Availability, Migration to Clouds and Compatibility, Interoperability and Communication between Clouds. In this paper, we concentrates on load balancing in cloud computing. We have considered artificial intelligence based algorithms for the cloud load balancing.

Key Words: Cloud Computing, Load Balancing & Artificial Intelligence

1. Introduction:

Cloud Computing can be defined by various form but the widely accepted definition, including by cloud security alliance [1] is given by NIST, that defines cloud computing as "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

The cloud computing exhibits, remarkable potential to provide cost effective, easy to manage, elastic, and powerful resources on the fly, over the Internet. The cloud computing, upsurges the capabilities of the hardware resources by optimal and shared utilization. The above mentioned features encourage the organizations and individual users to shift their applications and services to the cloud. Even the critical infrastructure, for example, power generation and distribution plants are being migrated to the cloud computing paradigm.

The cloud computing provides virtualized resources to the customers using various technologies, for example, Web services, virtualization, and multi-tenancy. The cloud services are delivered to the customer through the Internet [2]. The Web applications are used to access and manage cloud resources that makes Web applications an important component of the cloud computing [3]. The customers' processes are executed in virtualized environment that in turn utilize the physical resources [4]. Multiple virtual processes of various users are allocated to same physical machines that are segregated logically. This gives rise to a multi-tenant environment in the cloud. Despite the provided advantages, the cloud computing is not exclusive of risks with security and increasing cloud load being the key risk [5].

Heavy Load on cloud is one of these bigger retardation and load balancing becomes the biggest issue for the cloud computing. For efficient load balancing there

must be some parameters to evaluate the load balancing techniques to get better resource distribution for the user demands. Measurement parameters allow us to see whether the given technique is good enough to balance the load of the traffic on the server or not. These parameters include throughput, fault tolerance, response time, performance, scalability and resource utilization.

In this paper we are presenting a comparative review on the existing artificial intelligence based technique to balance the load cloud server by proper allocation of resources. This section presents the brief introduction of the concept. Rest of the paper is organized as below.

Section II presents the Cloud Computing Model. Section III explains the concept of load balancing. Section IV gives the comparative review of soft computing based load balancing concepts. Section V concludes he paper.

2. Cloud Computing Challenges:

In overall, cloud-based challenges have been classified to six main categories: Data Management and Resource Allocation, Security and Privacy, Load Balancing, Scalability and Availability, Migration to Clouds and Compatibility, Interoperability and Communication between Clouds. Each of these issues has affected the reliability and efficiency of cloud based environments regarding to their concepts.

A. Data Management and Resource Allocation:

One of the most important concepts in cloud computing data centers is resource allocation. The importance of this concept is specified according to the large number of resources in cloud-based environments. Accordingly, resource allocation process should fulfill network quality of service requirements, eliminate performance hiccups without significant enhancement of service provider cost, and manage energy consumption [6]. Challenges in resource allocations are classified into three major parts: data center network resources, data center processing resources, and energy efficient data center resource allocation.

B. Security and Privacy:

One of the most challenging issues that decrease the rate of reliability and efficiency in cloud computing environments is ensuring the security and privacy of stored resources. Cloud computing security has become an important topic in industry and academic research and has become the leading cause of impeding its development [7].

Typically, security issues in cloud-based environments have been divided to three main parts: vulnerability to attack, standard security practices and being subject to state or national data-storage laws related to privacy or record keeping. These issues led to the appearance of considerable concerns in various levels (*i.e.* service providers, infrastructure, and endusers).

C. Load Balancing:

Load balancing is an important issue in cloud computing environment that is related to storage utilization and download performance. The main objective in this case is to establish an algorithm for assigning tasks to the cloud nodes effectively according to exiting limitations (*e.g.* high communication delays and heterogeneity). Typically, challenges and issues related to load balancing in cloud computing environments are classified to four major parts: spatial distribution of the cloud nodes, data replication, performance, and point of failure.

D. Scalability and Availability:

The ability of adapting cloud capacity to on-demand services during occurrence of various workloads (*i.e.* static, periodic, once-in-a-lifetime, unpredictable, and

continuously changing workloads) is a challenging issue in offering cloud based services. The lack of this ability may cause performance degradation (in the peak of workload) or even over sizing (in the bottom of workload) during the provision of on-demand services.

E. Migration to Clouds and Compatibility:

The rapid growth and the popularity of cloud computing between users and enterprises has encouraged traditional IT providers to move and adapt their products (*e.g.* traditional applications, operating systems, middleware, etc.) to cloud based environments. However, the possibility of success in this migration process is a challenging issue due to the existing limitations in traditional IT products. Typically, an IT product should contain five main specifications to enhance the rate of success in migration process like Modularity, Portability, Changeability, Scalability and Backward Compatibility.

F. Interoperability and Communication Between Clouds:

The lack of interoperability between various cloud vendors is an important issue that has happened due to disparate approaches and structure between them. This deficiency may happen in different level of cloud based environments (*e.g.* when an infrastructure-as-a-service environment cannot be moved to any platform-as-a-service provider effortlessly) or in a same level between providers (*e.g.* when an Amazon tenant cannot simply move his resources to Force.com).

3. Load Balancing:

Load balancing is the process of improving the performance of the system by shifting of workload among the processors. Workload of a machine means the total processing time it requires to execute all the tasks assigned to the machine [8]. Load balancing is done so that every virtual machine in the cloud system does the same amount of work throughout therefore increasing the throughput and minimizing the response time. Figure 1 shows a typical structure of load balancing concept. Typically, challenges and issues related to load balancing in cloud computing environments are classified to four major parts: spatial distribution of the cloud nodes, data replication, performance, and point of failure.

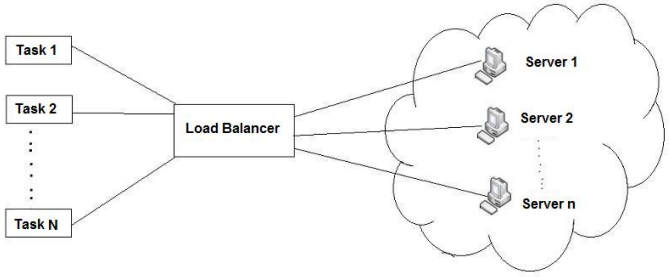

Figure 1: Load balancing

The first issue is the way of managing load balancing process between all spatially distributed cloud nodes. This management should consider several delays that can possibly happen because of the distance between clients and task processing nodes, speed of the network links between cloud nodes, and the distance between service nodes.

The second problem is the way the data is replicated (*i.e.* full and partial) in several nodes according to complexity of load balancing algorithm in various clouds nodes during partial replication and the additional requirement for more storage in full replication. The next issue is related to the performance of load balancing algorithm.

Due to the nature of load balancing processes, the complexity of the designed algorithm should be as less as possible to avoid faults and delays during intricate processes. The last challenge is the ability of resistance in load balancing algorithms against usual or unpredictable failures. Using a controller function to manage and minimize failures due to the increasing complexity of load balancing algorithms is a challenging issue in cloud based environments that attracts researchers' considerations.

4. Literature Survey:

Shojafar et al. (2015) has proposed an integrated approach of fuzzy logic and genetic algorithm name by 'FUGE' for the load balancing in cloud computing. Author has modified the standard genetic algorithm (SGA) and use fuzzy theory to devise a fuzzy-based steady-state GA in order to improve SGA performance in term of *makespan*. In the FUGE approach, author defines two types of chromosomes with different QoS parameters. Then, with the aid of fuzzy theory, authors obtain the fitness values for all chromosomes of the two types. After that, authors perform a modified fuzzy-based crossover operation on the two best selected chromosomes. The resulting chromosome is the output of the first generation; it will be added back into the population and a new population is created. In details, the *FUGE* algorithm assigns jobs to resources by considering virtual machine (VM) processing speed, VM memory, VM bandwidth, and the job lengths. The results of the experiments show the efficiency of the FUGE approach in terms of execution time, execution cost, and average degree of imbalance [9].

Issawi et al. (2015) has proposed burstness-aware load balancing algorithms in cloud environment. Here, Round-Robin approach is considered in burst state, Random in non-burst state and Fuzzy logic to assign the received request to a balanced Virtual machine. Cloud Analyst is used for the simulation of the results. Cloud Analyst is a graphical simulation tool based on Cloudsim for modeling and analyzing the behavior of a cloud computing environment, which supports visual modeling and simulation of large-scale applications that are deployed on Cloud Infrastructures. The experimental results shows the improved performance and decreased response & processing time as compare the other adaptive algorithms [10].

Wen et al. (2015) has proposed VM migration strategy based on Ant Colony Optimization for cloud computing load balancing. In this approach, local migration ants monitor the resource utilization and adapt two different traversing strategies to find the near optimal mapping between the virtual machines and physical machines. For the experimental evaluation of the load balancing, author has used the CloudSim toolkit package and shows the outperform migration results for the proposed ACO-VMM [11].

Mesbahi et al. (2014) has proposed a new cloud light weight model to balance the cloud load. In this algorithm, CloudSim cloud system simulator is used for the validation of algorithm. This algorithm balances the system load among all processing nodes in a cloud datacenter. Using this algorithm in our simulation, we balanced the cloud so that all its nodes have approximately the same weight in terms of distributing system workload. The main advantage of using algorithm is that it not only balances the cloud load but also gives assurance for the Quality of Services (QoS) for end users. It also reduces the migration time during execution and number of VM (Virtual machine) migration processes [12].

Florence & Shanthi (2014) have developed a firefly algorithm for load balancing in cloud computing. By using this algorithm load will be balanced properly. The proposed algorithm always gives the optimized result. The performance analysis of the proposed approach produced expected results so it is efficient for load balancing. In the

proposed approach CPU rate is efficiently utilized than the existing approach and load is properly balanced [13].

Kim et al. (2013) proposed a competent BABC (binary artificial bee colony) that contains a FRS (flexible ranking strategy) to get better balance between searching and utilization. For reducing the make span two different variants are established. Proposed algorithm is enhanced than some other approaches like GA (genetic algorithm), PSO (particle swarm optimization) and simulated annealing. For improving the result of swarm based technique and load balancing ant colony optimization technique is employed [14].

Dasgupta et al. (2013) have proposed a Genetic Algorithm (GA) over load balancing strategy. This optimization technique balances the load of the cloud environment, minimizing makespan of task. Cloud analyst simulator is used to simulate the proposed load balancing strategy. Some existing approaches like First Come First Serve (FCFS), Round Robin (RR) and Stochastic Hill Climbing (SHC) are used for simulation. GA utilizes all the dedicated resources associated with it. This proposed technique is better than few existing techniques; it also supports the QoS requirement [15].

5. Conclusions:

Cloud computing refers to the delivery of computing resources over the Internet. Instead of keeping data on your own hard drive or updating applications for your needs, you use a service over the Internet, at another location, to store your information or use its applications. But this increases the cloud load as the people have increased the use of cloud server. The problem has risen up to balance the cloud load.

In this paper, we have compared the artificial intelligence based load balancing techniques for the cloud server tasks. Different concepts of load balancing like fuzzy genetic approach, burstness aware load balancing, Ant colony based load balancing, BABC algorithm, Firefly based load balancing etc. concepts are considered. But still there is the need of more efficient approach to attain the proper balancing of cloud load as per increasing usage.

6. References:

1. Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." Communications of the ACM 53, no. 6 (2010): 50.
2. Duan, Qiang, Yuhong Yan, and Athanasios V. Vasilakos. "A survey on service-oriented network virtualization toward convergence of networking and cloud computing." Network and Service Management, IEEE Transactions on9, no. 4 (2012): 373-392.
3. Menzel, Michael, Lizhe Wang, Samee U. Khan, and Jinjun Chen. "CloudGenius: a hybrid decision support method for automating the migration of web application clusters to public clouds." Computers, IEEE Transactions on 64, no. 5 (2015): 1336-1348.
4. Guan, Bei, Jingzheng Wu, Yongji Wang, and Samee U. Khan. "Civsched: A communication-aware inter-vm scheduling technique for decreased network latency between co-located vms." Cloud Computing, IEEE Transactions on 2, no. 3 (2014): 320-332.
5. Latif, Rabia, Haider Abbas, Saïd Assar, and Qasim Ali. "Cloud computing risk assessment: a systematic literature review." In Future Information Technology, pp. 285-295. Springer Berlin Heidelberg, 2014.

6. Abu Sharkh, Mohamed, Manar Jammal, Abdallah Shami, and Abdelkader Ouda. "Resource allocation in a network-based cloud computing environment: design challenges." Communications Magazine, IEEE 51, no. 11 (2013): 46-52.
7. De Chaves, Shirlei A., Carlos B. Westphall, Carla M. Westphall, and Guilherme A. Gerônimo. "Customer security concerns in cloud computing." InProc. The Tenth International Conf. on Networks (ICN), The Netherlands, pp. 7-11. 2011.
8. Kherani, Foram F., and Jignesh Vania. "Load Balancing in cloud computing." (2014).
9. Shojafar, Mohammad, Saeed Javanmardi, Saeid Abolfazli, and Nicola Cordeschi. "FUGE: A joint meta-heuristic approach to cloud job scheduling algorithm using fuzzy theory and a genetic method." Cluster Computing 18, no. 2 (2015): 829-844.
10. Issawi, Sally F., Alaa Al Halees, and Mohammed Radi. "An Efficient Adaptive Load Balancing Algorithm for Cloud Computing Under Bursty Workloads."Engineering, Technology & Applied Science Research 5, no. 3 (2015): pp-795.
11. Wen, Wei-Tao, Chang-Dong Wang, De-Shen Wu, and Ying-Yan Xie. "An ACO-Based Scheduling Strategy on Load Balancing in Cloud Computing Environment." In Frontier of Computer Science and Technology (FCST), 2015 Ninth International Conference on, pp. 364-369. IEEE, 2015.
12. Mesbahi, Mehran, Amir Masoud Rahmani, and Anthony Theodore Chronopoulos. "Cloud light weight: A new solution for load balancing in cloud computing." In Data Science & Engineering (ICDSE), 2014 International Conference on, pp. 44-50. IEEE, 2014.
13. Florence, A. Paulin, and V. Shanthi. "A load balancing model using firefly algorithm in cloud computing." Journal of Computer Science 10, no. 7 (2014): 1156.
14. Kim, Sung-Soo, Ji-Hwan Byeon, Hongbo Liu, Ajith Abraham, and Sean McLoone. "Optimal job scheduling in grid computing using efficient binary artificial bee colony optimization." Soft Computing 17, no. 5 (2013): 867-882.
15. Dasgupta, Kousik, Brototi Mandal, Paramartha Dutta, Jyotsna Kumar Mandal, and Santanu Dam. "A genetic algorithm (GA) based load balancing strategy for cloud computing." Procedia Technology 10 (2013): 340-347.