International Journal of Innovations & Research Analysis (IJIRA)
ISSN :2583-0295, Impact Factor: 5.449, Volume 03, No. 01(II), January- March, 2023, pp 69-72

STUDY OF LOAD BALANCING FRAMEWORK IN CLOUD COMPUTING

Sonam Jain*

ABSTRACT

Cloud computing is a technology that is based on virtualization allows us to create, configure, and customize applications via an internet connection. Cloud computing is distributed computing, storing, sharing and accessing data over the internet. It provides lots of shared resources to its users on per day use basis meaning users are only billed for services they use according to their access times. The amount of data being processed and kept in cloud environments is increasing very rapidly. Cloud computing is generating a lot of data processing and storage every day. This leads to an uneven distribution of workload across cloud resources. The load balancing techniques is defined as dividing workload and computing properties in cloud computing. Cloud computing already has different scheduling algorithms, including Shortest Job First (SJF), First Come First Serve (FCFS), Round Robin (RR), etc. so by using Scheduling tasks efficiently has become an important problem to be solved in the field of cloud computing Among the many parameters to consider when load balancing in cloud computing, make span and response time are the most important. In order to solve the load balancing problem, this paper provides a summary of evolutionary and swarm-based algorithms that will help to solve the problem in different cloud environments.

Keywords: *Cloud Computing, Load Balancing, Swarm Based Algorithms, Quality of Services (QoS), Distributed Computing, ant Colony Optimization (ACO), Artificial Bee Colony Algorithm (ABC), Resource Allocation, Computation Energy, Virtual Machine (VM).*

Introduction

In recent years, cloud computing has become very popular [1,3]. There is an uneven and heavy workload on cloud resources as the size of computation and demand for higher computation is growing rapidly. All of the loads are distributed among all nodes by load balancing [4]. All of the processing units are evenly distributed. Basically, it prevents bottlenecks from forming in the overall system due to load disparity [5]. It is important to remember that load balancing is one of the issues associated with this field [6]. We can conclude that cloud computing provides virtual infrastructure backed by software and hardware resources on the internet. With cloud computing, a user has access to a shared pool of resources on demand, to which he subscribes and can access for as long as he wants, using virtualization to reduce the cost of implementing additional hardware to meet the needs of the user [2]. The deployment models of Cloud computing have been categorized into four categories. Today, three categories are most commonly used. A public cloud is also known as an open cloud model that is the most common model of cloud computing. With this model, cloud services are provided through a virtualized environment developed using polled shared physical resources shared over a public network such as the Internet. Multiple clients share the same infrastructure. With this model, cloud operations are optimized. A private cloud is designed and developed to fit the needs of a single organization. A private cloud service provider grants you access to its network in a more secure way, ensuring that anyone

* Lecturer, Department of Computer Science, S.S. Jain Subodh P.G. Mahila Mahavidyalaya, Jaipur, Rajasthan, India.

outside your network will not be able to access it. Cloud computing platforms are highly scalable. Depending on the needs of the user, it can be scaled up or down [7]. Due to cloud platforms' dynamic nature, they require efficient and effective load balancing across all machines to reduce makespan, the response time of a single task, energy consumption, and interruption of services. When load balancing is performed correctly across different cloud resources, then there is also high availability of services if any of the other resources aren't responding. A variety of task scheduling techniques are available in cloud computing. The main problem is that basic load balancing techniques don't utilize resources effectively. Therefore, it increases the overall processing time in cloud computing [5]. Cloud computing's virtualization capacity hides the diversity of resources that make it different from other technological advances.

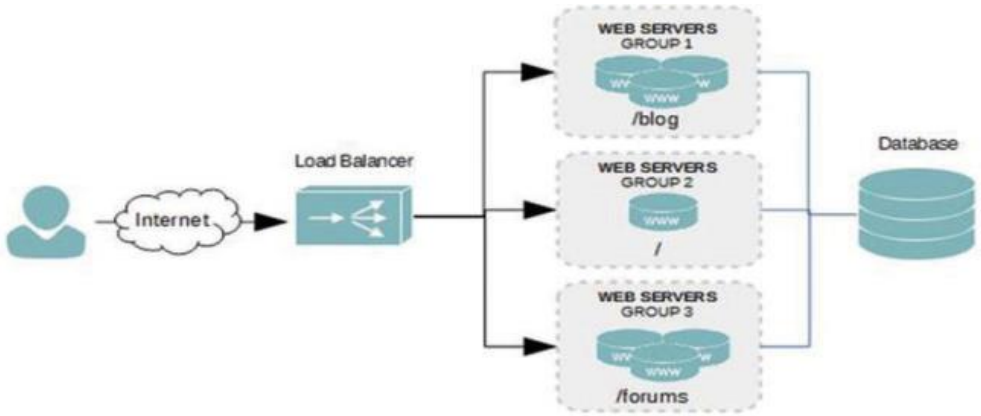

Fig. 1: Load Balancing Groups [2]

Related Work

Workload distribution in a balanced manner is a key challenge of cloud computing. By distributing workload across multiple nodes, resources are properly utilized. This is an optimization problem and a good load balancer should be used for this strategy due to the types of tasks and dynamic environment. [9]. For the load scheduling problem and simple applications, many better approaches are presented. Kousik Dasgupta et al. (2013) [9] have proposed a Genetic Algorithm (GA) over load balancing strategy. Using this optimization technique, the load on the cloud environment can be balanced, minimizing the time it takes to complete a task. Cloud analyst simulator is used to model the load balancing strategy. Simulations use methods such as First Come First Serve (FCFS), Round Robin (RR) and Stochastic Hill Climbing (SHC). GA uses all the dedicated resources associated with it. The proposed technique provides better QoS than few existing techniques. A. Paulin Florence et al. (2014) [10] have developed a firefly algorithm for load balancing in cloud computing. The load will be balanced correctly using this algorithm. This algorithm always gives the best results.In a cloud computing environment, load balancing refers to how requests are distributed among resources to betterresults. In previous load balancing algorithms, there have been many constraints. The FCFS algorithm is the simplest load balancing algorithm because it considers only the time it takes a task to reach the virtual machine. In Round Robin (RR), tasks are assigned the same time slots, whereas FCFS does not. Previous load balancing algorithms, such as Shortest Job First (SJF) and FCFS, did not improve throughput much.

A RR load balancer is one of the simplest and most common techniques used in cloud computing environments. It allocates tasks to different resources based on time units. Its load balancing technique ensures that all available cloud resources are utilized in RR fashion by randomly selecting any node and assigning it the first task. In cases where all nodes are occupied at the time of task allocation, the node with the smallest queue of tasks is assigned the task. The cut-off times for tasks different[1]. As RR allocates tasks to VMs, it does not take into account the length of the tasks, the processing power of the VM, or the priority of the tasks. Consequently, every node has equal weight during execution and will be allocated equally. Hence, author in states that RR algorithm is based on random sampling i.e. it selects load randomly. In this execution, some nodes are heavily loaded while some are lightly loaded with tasks. Hence RR does not optimize resource utilization. SheetalKarki et al. in 2018 explains that the data is kept in a centralized virtual machine called cloud and the cloud provider companies are

responsible to assign the offerings to the end users[5]. The end users get entry to the offerings primarily based on their needs and are to be paid for what's being served. As the number of requests grows so the need for load balancing arises to maximize the useful resource used and energy consumption. Threshold and Check Pointing algorithm help in task migration when the virtual machines get overloaded at the time of cloudlet execution. The tasks are moved from one virtual machine to another or can be queued to be decided by threshold and check pointing algorithm minimizing the processing time, energy and resource consumption. According to Mayur S. Pilavare et al. in 2015, cloud computing is connected to servers via a network, so there are many issues to be resolved. Load balancing is the important issue over the cloud to be addressed. The Genetic Algorithm outperforms some existing load balancing techniques. As a result of giving the prioritized input to the genetic algorithm, the response time will be decreased, minimizing the duration of the given task set. The jobs here are assumed to have the same priority but that may not actually be the case, so it can be taken for further analysis and the various selection techniques for GA can be changed for better performance, and crossover and variation techniques can be modified to get better performance [5].

Proposed Methodology

An improved genetic algorithm detects all free virtual machines within the system. When a new task arrives, the availability of a free virtual machine is checked. If a virtual machine is available, the task is assigned to it. In the absence of a virtual machine, the machine that will complete the first task is assigned the next[9]. As a result, all the VMs are utilised properly and there is no overutilization or idle situation here. In comparison to other techniques, the results gained here are more cost-effective and energy-efficient. Following are the various steps of Improved Genetic Algorithm:

Initialization

Normally, candidates are generated arbitrarily across the search area. However, domain-specific knowledge or other information can easily be incorporated.

Evaluation

When the population is initialized or an offspring population is created, the fitness values of the candidate solutions are evaluated.

Selection

Therefore, the candidates with better fitness values are given more copies of the selection and survival-of-the-fittest is imposed on them. This idea proposes choosing a better selection over a worse one, and multiple selection theories have been used to accomplish this, including stochastic universal selection. Other selections include roulette-wheel selection, ranking selection, and a few uncommon selections such as tournament selection, which are well described in the following report.

Recombination

It involves merging the components of two or more parental solutions to create a new solution (offspring) that may be better. A well designed recombination mechanism is crucial for competent overall performance. In recombination, the offspring will not be identical to any particular parent and will combine parental developments in a novel way. Implementation can be broken down into the following steps:

- **Step 1:** In the very first step, the cloud network is deployed with a finite number of virtual machines.
- **Step 2:** During the second step, the best virtual machine is chosen for the cloudlet execution, and when the fault occurs in the network the next step is executed.
- **Step 3:** The improved genetic algorithm is then executed, which reassigns the task to the other virtual machine when a fault occurs and continues the process.

Future Scope

The proposed algorithm can be further optimized by combining it with other virtual machine migration algorithms and enhancing it by utilizing a hybrid approach utilizing meta-heuristic algorithms in the future.

Conclusion

Due to its dynamic nature, cloud computing has a number of security, quality of service, and fault occurrence issues. Load balancing is the primary concern of cloud networks, which minimizes efficiency. An enhanced genetic algorithm is an algorithm applied to existing work when faults are detected in order to migrate virtual machines[2]. Load balancing in cloud computing can be implemented

in many different ways. Nevertheless, none of them has addressed all the load balancing issues. In the previous sections of this report, we explained that load balancing in cloud computing takes into account a variety of parameters and attempts to improve cloud system performance based on those parameters[6]. These parameters include the response time of machines to users' requests, the span of VMs, and resource optimization, among others. Cloud resource load balancing is a useful tool for sharing computing workloads across multiple regions by maximizing efficiency of cloud resources. For better computing load, the automatic scaling listener monitored network traffic and distributed dynamic load equally across multiple cross-regions[5]. To improve access time in different regions, we will explore efficient load balancing algorithms that can find IP addresses through API hubs. For restaurants revenue method, the use of Elastic Load Balancing minimizes the operational overhead and monitors the network traffic for the different domain-oriented restaurants for many purposes, including order taking and revenue sharing[3].

References

1. Srinivas.J, K. Venkata Subba Reddy, “Cloud Computing Basics”, International journal of advanced research in computer and communication engineering, 2012, pp. 343-347.
2. Soumya Ray and Ajanta Sarkar, “Execution Analysis of Load Balancing Algorithm in Cloud computing Environment”, International Journal on Cloud Computing: Services and Architecture (IJCCSA), October 2012, Vol.2, No.5.
3. HU Baofang, SUN Xiuli, LI Ying, SUN Hongfeng, “An Improved Adaptive Genetic Algorithm in Cloud Computing”, 13th International Conference on Parallel and Distributed Computing, Applications and Technologies, 2012.
4. Rajiv Ranjan, Liang Zhao, Andres Quiroz and Manish Parashar, "Peer-to-Peer Cloud Provisioning: Service Discovery and Load Balancing", Cloud computing: Principles, Systems and Applications, computer communications and networks, Springer, 2010, pp. 195-217.
5. Borja Sotomayor, Ruben S.Montero, Ignacio M. Llorente, and Ian Foster, "Virtual infrastructure management in private and hybrid clouds", IEEE Internet Computing, 2009, pp. 14-22 .
6. Ali M. Alakeel" A guide to dynamic load balancing in distributed computer systems", International Journal of Computer Science and Network Security, VOL.10 No.6, 2010 ,153-160.
7. ThilinaGunarathne, Tak-Lon Wu, Judy Qiu and Geoffrey Fox, "MapReduce in the Clouds for Science" 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010, pp.565-572.
8. A. Nuaimi, A. Jaroodi, “A survey of load balancing in cloud computing: challenges and algorithm”, in Proc - IEEE 2nd Symp. Netw. Cloud Comput. Appl. NCCA, 2012, pp. 137–142.
9. X. Shao, Y. Teranishi, N. Nishinaga, “Effective load balancing mechanism for heterogeneous range queriable cloud storage”, presented at IEEE 7th Int. Conf. Cloud Computing Technology, 2015, pp. 405–412
10. H. Sun, T. Zhao, Y. Tang, “A QoS-aware load balancing policy in multi-tenancy environment”, Proc. - IEEE 8th International Symposium on Service Oriented Syst. Engineering, 2014, pp. 140–147.

□❍□