International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 3, Issue 1, pp 37-41, January-2019

Approach for Real-Time Scaling in Load Balancing for Effective Resource Utilization in Cloud Computing

Gopaji Monica

Assistant Professor, Department of Computer Science And Engineering, Malla Reddy College of Engineering for Women, Hyderabad -500100, Telangana, India

Abstract - Cloud computing is a current model for getting to administrations by denotes of the web. This model has a few strains, for example, load-adjusting, security measures, asset orchestrating scaling, Quality of Service (QoS) control, administration availability, and server farm energy use. Among these, quite possibly the most prominent trials are load-balancing. Load balancing issue is a multivariate, multi-requisite issue that corrupts the execution and efficacy of processing assets. Load balancing methods correspond with the answer for load imbalanced circumstances for two unwanted aspects of overloading and under-stacking. Load balancing minimizes the overhead and maximizes throughput by dividing the tasks among the available machines using various suitable load balancing algorithms. In this paper, we have provided an overview of various aspects of load balancing and its algorithms.

Keywords: cloud computing, load balancing, significance, SWOT analysis, goals, round robin, stochastic hill climbing, max-min, resource allocation.

1. INTRODUCTION

Cloud computing is a web-based advancement or a network technology based on the internet that has a part in the swift progress of communication technology, giving a platform for applications and services and a way to configure and adjust. It is a decentralized way of computing with location independence, device independence computational process. Cloud is usually referred to as 'ubiquitous' which means 'being present everywhere at the same time.' and its contents are configurable and shareable. It has led to the advancement of distributed systems to an extensive computing network using which, firms like Amazon, IBM, Google & Yahoo deliver cloud services to users all over the world. Here, apps and services are offered on-demand to end-users and hence, they need not install it on their local systems. Load Balancing insinuates the distribution of the inevitable load among various computer collections, computer solutions, relation to the network, disks, servers, CPUs, etc. ensuring that no computing machine is under-loaded, overloaded or idle. It helps in preventing the occurrence of a deadlock and overloading, and assists networks and resources by providing a high throughput and minimum response time and is a major challenge faced in cloud computing.

LB proposes ways to maximize the system output, device performance, usage of resources and also offers accessibility, scalability and availability. The efficiency of a cloud computing model is determined by its utilization of the resources. The best results can be attained by implementing and properly managing the cloud resources. These resources are given to the users through VMs that are Virtual Machines. They make use of Virtualization which utilizes hardware, software or an entity called a hypervisor. Our main aim is to analyse the types of load balancing and its different algorithms along with a comparative study of their advantages and drawbacks, in this paper. The paper also elaborates on the significance of load balancing and the challenges faced along with suggested methods that can be used in the future.

2. SIGNIFICANCE OF LOAD BALANCING ALGORITHMS

Load Balancing is helpful in cloud environments where immense workloads overwhelm a server easily. As certain performance metrics like availability of service and response time become crucial to some business operations, the need for load balancing also increases. Load balancing is a way to identify available servers and redirect the traffic to them while one server is being overloaded. This ensures no server is sitting idle. Thus, if Load Balancing is not ensured, the new virtual servers won't be able to manage the incoming traffic in an organised manner.

Further, a few benefits of cloud computing are listed using the SWOT analysis (Strength, Weakness, Opportunities and Threat Analysis.) [1]

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 3, Issue 1, pp 37-41, January-2019

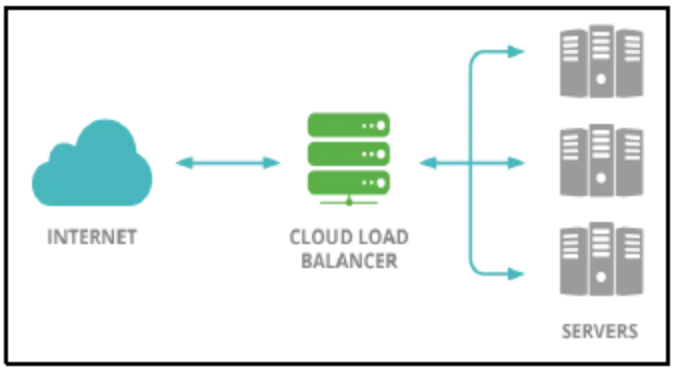

Fig -1: Load Balancing in Cloud Computing

3. TYPES OF LOAD BALANCING

There are majorly two categories of load balancing based on the virtual machine's current state- static and dynamic. These are as follows:

Static Load Balancing

In static load balancing [3], the information and data about the system like the processing power, storage requirements and client necessities is known beforehand. The condition of the system is already known such as the job resource requirements, processing power of the system, time of computation and the size of the memory and storage device. It follows a collection of predefined rules which don't need to know the current state of the network. This strategy is not extensible although it is quick and efficient and hence it operates well when there is low load variance in the nodes. Its operating period is less as compared to that of dynamic load balancing. Uncertain resource distribution is caused due to the failure of finding the connected servers. A major disadvantage of static load balancing is that the system's actual state is given very less importance to decision making and hence distributed systems with constantly changing states are unacceptable. It does not consider continuous monitoring of the nodes and hence cannot consider load changes during run-time. The strategies for static load-balancing are [4]:

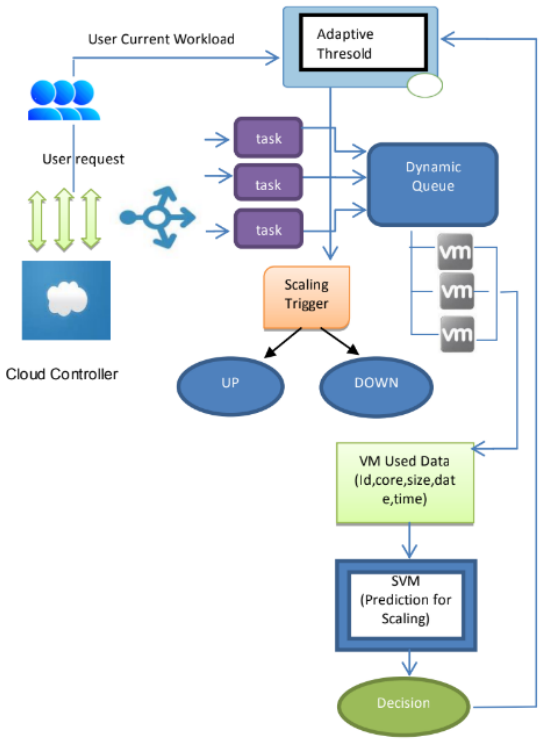

Fig -2: Static Load Balancing

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 3, Issue 1, pp 37-41, January-2019

a. Optimal: Resource information is gathered by the data communication network using structured techniques which is sent to the load balancer where maximum allocation is performed in a limited time period.

b. Sub Optimal: If a decision cannot be correctly given by a load balancer, a suboptimal solution will be decided instead. Min-Min load balancing, Max-Min load balancing, Round Robin, Shortest Job First, Throttled load balancing, Two-phase Opportunistic load balancing, and Central LB are just some static load balancing algorithms.

Dynamic Load Balancing

This approach [3] is more accurate and efficient and it considers the current state of the system to make further decisions. Dynamic load balancing leverages the fact that it allows the tasks to transfer to an underloaded machine from an overloaded machine and therefore it is adaptable which leads to an improved performance. Some other benefits of dynamic load balancing include increased scalability, resilience to faults, and reduced expenses to increase efficiency which can also handle unreliable processor loads. It monitors the loading of nodes while processing, regularly to calculate the node workload and redistribute the workload among the nodes. Although dynamic load balancing approaches are adaptive in nature and are good for fault tolerance, they have less stability and have high utilization of resources.

Challenges of Load Balancing

The most pressing challenges of load balancing [4] are as listed below:

1. Distributed Geographical Nodes
2. Single Point of Failure
3. VM Migration
4. Heterogeneous Nodes
5. Handling Data
6. Load Balancer Scalability
7. Algorithm Complexity
8. Automated Service Provisioning [6]

4. ALGORITHMS OF LOAD BALANCING

Round Robin Algorithm

Introduction:

1. Round Robin is an algorithm for Static load balancing that is it depends on the past knowledge of resources and software of the system and the choice of distributing workload does not entirely depend on the system's present status.

2. It uses Round robin fashion to allocate jobs and this scheduling is an effective and efficient time scheduling policy.

3. The nodes for load balancing are randomly selected by this algorithm.

4. The important duty, here, is the process of handling load balancing in cloud computing which is carried out by data centers.

5. When the controllers of the data centers get requests from the user, then this request is passed to the round robin algorithm.

6. The partitioning of time in parts inside the RR algorithm is called time quantum or slices of time. Hence, this algorithm is uniquely made for dividing time.[2]

Method:

1. Initially, all processors are kept in a queue that is circular.

2. For every processor within the queue, the server is allocated by the scheduler in the defined slot of time.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 3, Issue 1, pp 37-41, January-2019

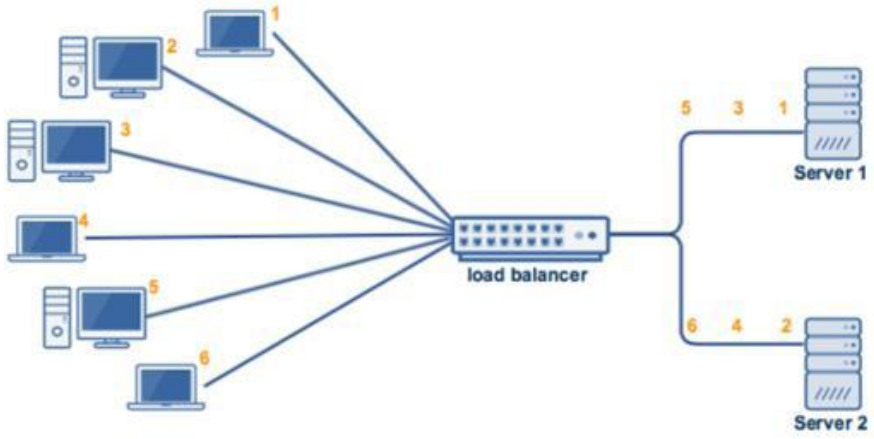

Fig -3: Round Robin Algorithm

BT: (Burst time) The execution period required by a process.

TQ: (Time Quantum) The amount of time allowed for a processor to run.

In Round Robin Algorithm, since the servers are selected on a random basis, there are chances that few servers may be loaded more than their capacity. This will in turn lead to the decline in the performance of load balancing. To overcome this issue, a weighted round robin load balancing algorithm is used which is an extended version of the round robin method. [2] In this method, the administrator can assign a weight to each server based on criteria like traffic handling capacity. So, the servers will gain more requests from clients since they are allocated with more weights.

5. RESULTS AND DISCUSSION

The above algorithms were compared based on their performance metrics as described in [12][13]. This approach [3] is more accurate and efficient and it considers the current state of the system to make further decisions. Dynamic load balancing leverages the fact that it allows the tasks to transfer to an under loaded machine from an overloaded machine and therefore it is adaptable which leads to an improved performance. Some other benefits of dynamic load balancing include increased scalability, resilience to faults, and reduced expenses to increase efficiency which can also handle unreliable processor loads. It monitors the loading of nodes while processing, regularly to calculate the node workload and redistribute the workload among the nodes. Although dynamic load balancing approaches are adaptive in nature and are good for fault tolerance, they have less stability and have high utilization of resources.

6. CONCLUSIONS AND FUTURE WORK

Cloud computing ensures the delivery of customer support at all times. A major challenge in Cloud computing is load balancing, since overloading of a device may lead to dreadful results. Hence, there is a constant need for an effective LB algorithm with the help of which resources can be efficiently utilized. The major aim of load balancing is to serve the user's requirements by allocating the work-load across several network nodes and amplifying the usage of resources, thereby increasing the performance of the cloud system, minimizing the response time, and reducing the number of job rejection which results in a reduction of the energy consumed and the carbon emission rate. This paper describes the significance of cloud computing, goals of Load balancing in cloud computing, types of load balancing and load balancing algorithms. We described five load balancing algorithms and presented their comparison based on their performance metrics. There will be demand for new fully autonomous dynamic Load Balancing algorithms in the future that will allow increased resource utilization, improved degree of mismatch, effective task migrations, lowered make-span and shorter time span.

REFERENCES

[1] R. Z. Khan and M. O. Ahmad, "Load balancing challenges in cloud computing: A survey", Proc. Int. Conf. Signal Netw. Comput. Syst., vol. 396, pp. 25-32, 2016.

ISSN (online): 2581-3048
Volume 3, Issue 1, pp 37-41, January-2019

[2] Brototi Mondala, Kousik Dasguptaa, Paramartha Dutta: "Load Balancing in Cloud Computing using Stochastic Hill Climbing-A Soft Computing Approach", Procedia Technology 4 (2012): 783 – 789

[3] Kanani, Bhavisha, and Bhumi Maniyar. "Review on max-min task scheduling algorithm for cloud computing." Journal of Emerging Technologies and Innovative Research 2.3 (2015): 781-784.

[4] V. Soni and Dr. N.C. Barwar: "Performance Analysis of Enhanced Max-Min and Min-Min Task Scheduling Algorithms in Cloud Computing Environment" in International Conference on Emerging Trends in Science, Engineering and Management (ICETSEM-2018), Pune, Maharashtra, India on 14th October 2018

[5] V. Joshi and U. Thakkar, "A Novel Approach for Real-Time Scaling in Load Balancing for Effective Resource Utilization," 2018 International Conference on Smart City and Emerging Technology (ICSCET), Mumbai, India, 2018.

[6] S. Kumar Mishra, B. Sahoo and P. Paramita Parida :"Load balancing in cloud computing: A big picture", Journal of King Saud University – Computer and Information Sciences 32, (2020)

[7] U. Patel and H. Gupta, "A Review of Load Balancing Technique in Cloud Computing" Published online: April 24, 2019.

[8] S. Rao Gundu1, C. Arur Panem and A.Thimmapuram: "Real Time Cloud Based Load Balance Algorithms and an Analysis", Published online: 28 May 2020

[9] N.Tadapaneni, "A Survey of Various Load Balancing Algorithms in Cloud Computing" Published online: April 2020.

[10] N. Verma, V. Sharma, M. Kashyap and A. Jha, "Heuristic Load Balancing Algorithms in Vulnerable Cloud Computing Environment," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2018.

[11] Muhammad Asim Shahid, Noman Islam, Muhammad Mansoor Alam, Mazliham Mohd Su'ud and Shahrulniza Musa: "A Comprehensive Study of Load Balancing Approaches in the Cloud Computing Environment and a Novel Fault Tolerance Approach", July 27, 2020.

[12] P. Kumar and R. Kumar, "Issues and challenges of load balancing techniques in cloud computing: A survey", ACM Comput. Surveys, vol. 51, no. 6, pp. 1-35, Feb. 2019.

****

