International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 10, Issue 2, February 2021)

Load Balancing Techniques in Cloud Computing: A Review

Swati Agrahari[1], Sapna Choudhary[2]

[1]M.Tech Scholar, [2]Prof. Dept. of CSE, SRGI, Jabalpur, India

swatiagrahari25@gmail.com, choudharysapnajain@gmail.com

***Abstract*-- Cloud computing is a rapidly expanding field of computing science and industry today, with numerous advantages including reduced time, limitless computing capacity, and versatile computing capabilities. It's a model that allows users to connect to a common pool of computing resources on demand. It encompasses a wide range of principles, including load balancing, scheduling, and so on. From the convergence and evolution of industry, technology, and methodology perspectives, cloud computing as computing resources developed as a new kind of service in the field of communication and information technology that enables and furnishes its users to access IT resources anywhere and anytime at low cost on a pay-per-use basis. The resources over cloud must be handled and controlled by an appropriate mechanism by setting the priorities on the available cloud resources which are shared among its users. To overcome these challenges, this paper conducts a systematic literature review on load balancing algorithms and challenges in cloud computing for one of the most important cloud resources, the Virtual Machine over Cloud environment.**

***Keywords* – Cloud, Virtual Machine, Distributed, Load Balancing, Migration, Scheduling**

I. INTRODUCTION

Cloud computing is a relatively new technology that has gained a lot of traction in recent years. It allows users to rent software, hardware, infrastructure, and computational resources on a per-user basis. Prior to the development of cloud computing systems, there were Client-Server programmes, which ran on a server and had centralized storage, as well as all of the system's controls. If a user wants to retrieve information from the server or deploy an application on the server, he or she must first obtain sufficient access and then log in to do so. The number of cloud users is rapidly increasing, and it appears that virtual machine scheduling in the cloud is becoming an important problem to investigate. Users may either send their jobs to the cloud for statistical processing or store their data there. The job must be correctly scheduled by the cloud scheduler [1].

It is an entirely internet-based approach in which all programmes and data are hosted on a cloud that is made up of thousands of computers that are intricately connected together. There are new distributed networks that operate on a pay-per-use basis [2]. Load balancing is use to balance load between multiple resources to get minimum makespan, improve performance, reduce response time and optimal resource utilization. Cloud computing has many advantages which makes any industry to move from conventional infrastructure. Figure 1 show why we should move to cloud computing and adopt cloud computing infrastructure.

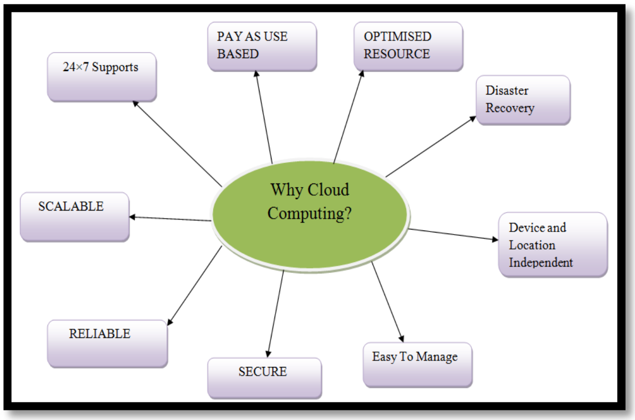

Fig. 1: Characteristics of Cloud Computing

1.1 Load Balancing:

LB's primary objective is to efficiently manage the load across various cloud nodes, so that no node is under / overloaded [7]. LB may be characterized as a process of spreading a burden across network links on multiple devices or system clusters to maximize its use of assets to optimize overall response time. It reduces the device's total waiting period and also avoids excessive replication of assets. Requests spread inside servers in this process so that data can be distributed & processed without waiting. LB is the method of maximizing system performance by moving the device burden [9]. The LB at CC is shown in Figure 4.

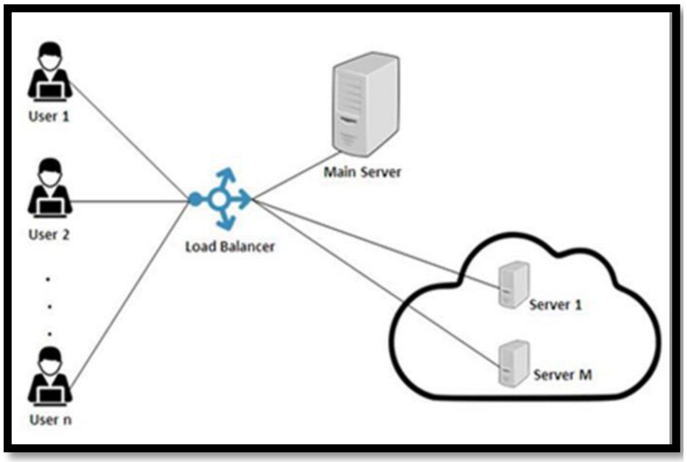

Fig. 2: Load balancing in Cloud Computing

LB provides a systematic mechanism for the equal distribution of the responsibility to the resources available. The goal is to provide reliable service, including adequate use of the resource, in the event of a disaster of the portion of any service by supplying & de-provisioning the device instance. In addition, LB is aimed at reducing response time for tasks & increasing resource efficiency, which increases device efficiency at a lower cost [9].

This paper has been divided into the following sections. Section 2 specifies the literature study relevant to this work. Section 3 explains the existing load balancing techniques. Section 4 discusses the challenges in cloud computing. Section 5 deals with the conclusion and future directions.

II. RELATED WORK

Load balancing is a critical problem in cloud computing. Many methods to dealing with load balancing problems in cloud computing systems have been suggested. All of these efforts seem to be aimed at improving the process of distributing workload across cloud nodes in order to achieve optimal resource usage, minimum data processing time, and minimum average response time, as well as avoid overloading.

The distributed dynamic priority based algorithm is used to efficiently balance the load on instances, improve system consistency, reduce response time, and increase throughput. Prioritizing resource allocation on virtual machines results in faster response and processing times. Load balancing ensures that all instances in a network node are doing the same amount of work at any given time. Priority-based resource provisioning to increase resource efficiency and reduce cloud service response times. [1]

Kulkarni A.K. (2015) proposed a variant of active VM algorithm to solve the issue during peak hours by using a Reservation table. The proposed VM load balancer keeps an internal reservation table to keep track of VM reservations suggested by the load balancer to the data centre controller, but this information is not changed in the allocation table before the allocation notification arrives. The proposed load balancer considers all reservations table entries and allocation statistics table entries for a specific VM id when selecting a VM for the next order. In this paper, an effective VM load balancing algorithm is proposed that distributes the load equally across all VMs in the data centre, even during peak hours when the incoming request frequency is high [3].

James J. et al. (2012) proposed Weighted Active Monitoring Algorithm is a modified variant of the Active Monitoring Algorithm that assigns weight to each node in order to improve response time and processing time. The VMs are allocated varying (different) amounts of the available processing power of the server/physical host to the individual application services in this proposed Load balancing algorithm using the principle of weights in active monitoring. Tasks/requests (application services) are delegated or allocated to these VMs with varying processing forces, starting with the most powerful, then the least powerful, and so on, based on their weight and availability. As a result, optimising the given performance parameter is important [4].

Around the year 1961, John MacCharty [5] suggested in a speech at MIT that “Computing can be sold like a utility, just like a water or electricity”. This brilliant and innovative idea was much ahead of its time, but fortunately this became a realistic one after few years.

A new Priority based job scheduling algorithm (PJSC) in cloud computing. This algorithm is based on multiple criteria decision making model and mathematical model called Analytical Hierarchy Process (AHP) provides scheduling with minimum makespan, high throughput and reasonable complexity.[6]

III. EXISTING LOAD BALANCING TECHNIQUES IN CLOUD COMPUTING

Different load balancing methods have already been applied in cloud computing using tools such as Cloud Analyst and Cloud-Sim. The following is a brief summary of the methods used:

Active Monitoring Load Balancing (AMLB) Algorithm: It keeps track of each VM's current workload and the number of requests assigned to it. When a request for a new VM is sent, the least loaded VM is identified. If there are many, the first one found is chosen. The Data Centre Controller receives the VM id from the Load Balancer. It sends the request to the VM with that id, and the Active VM Load Balancer is notified of the new allocation. Only the current load of the VM is taken into account when allocating it, and its computing capacity is ignored. As a result, the time it takes for certain jobs to be completed can increase, contravening the QOS requirement. [7]

Round Robin: Round robin is a simple and static scheduling strategy that works on the theory of time slices, which divides time into multiple intervals and assigns each VM to one of them. Is given a specific time slice or interval to work with. Round robin uses a random array of virtual machines. It rotates requests among a list of available virtual machines. The first request is allocated to a VM chosen at random from the party, and the requests are then distributed in a circular order by the Data Centre controller. When a VM is allocated to a request, it is moved to the bottom of the queue [8].

Throttled Algorithm: The load balancer keeps an index table of virtual machines and their states here (Available or Busy). The client/server first requests that the data centre locate a suitable virtual machine (VM) to perform the recommended task. The data centre asks the load balancer to assign the VM to it. The load balancer scans the index table from top to bottom until it finds the first available VM or the index table is fully scanned. If the VM is located, the load data centre is selected. The request is sent to the VM specified by the id by the data centre. Furthermore, the data centre notifies the load balancer of the new allocation, and the index table is updated accordingly [9].

Equally spread current execution load: This algorithm necessitates the use of a load balancer to keep track of the jobs that are being executed. The load balancer's function is to queue jobs and distribute them to various virtual machines. The balancer scans the queue for new jobs on a regular basis and assigns them to the list of available virtual servers. The balance also keeps track of the tasks assigned to virtual servers, allowing them to see which virtual machines are idle and need to be assigned new tasks.

The cloud analyst simulation is used to carry out the experimental work for this algorithm. As the name implies, this algorithm works by evenly distributing the execution load across many virtual machines.

The diagrammatical representation of the algorithm used for load balancing in a cloud computing setting is shown in the following figure [8], [9]. The figure also depicts the three algorithms studied in this paper using the cloud analyst simulation tool, which is based on cloud sim and offers a graphical user interface for performing the experiments.

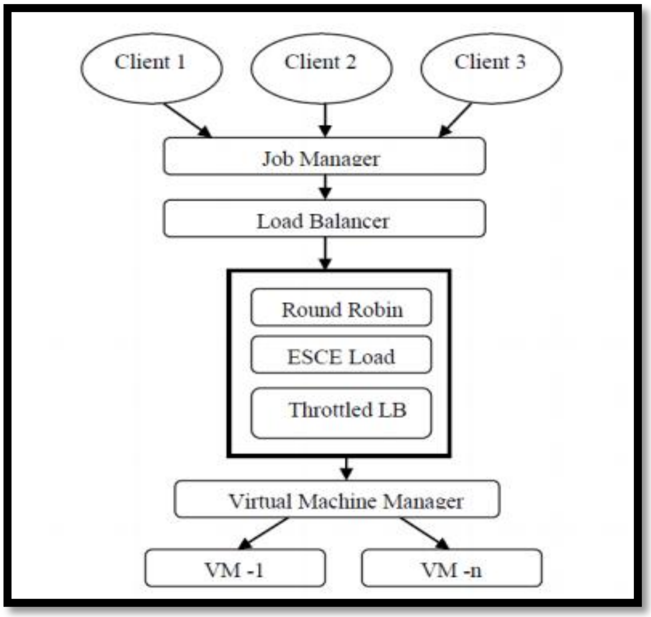

Fig.3: Load Balancing Algorithms Execution

IV. THE CHALLENGES OF CLOUD COMPUTING

There has been a lot of challenges associated with CC. Figure 3 demonstrates the taxonomy of major problems with the CC [5]. This includes: data protection [5], data recovery and availability [5], administrative capabilities [5], regulation and compliance restrictions [5], security [6], capable of adjusting the burden [6], controlling executions [6], load balance, fault tolerance, cloud computing governance, interoperability and portability .

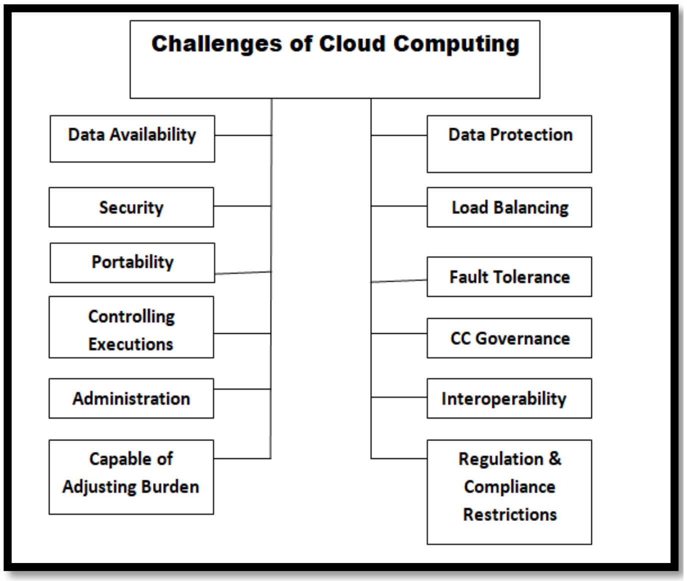

Fig. 4: Taxonomy of major problems in Cloud Computing

V. CONCLUSION

This article is focused on cloud computing problems and its major challenges. Cloud computing is state-of-the-art computer technology which delivers customer support at all times. Load Balancing is one of the biggest problems with Cloud Computing, as overloading a device will lead to terrible results that could create technology obsolete. So there is always a need for an effective LB algorithm for efficient use of resources. The main goal of LB is to meet user needs by distributing the workload across multiple network nodes & maximizing resource usage & growing device efficiency. The present paper also discusses the challenges associated with cloud computing.

Consequently, effective load management is critical for system efficiency, resource usage, reliability, throughput optimization and response time minimization.

REFERENCES

[1] G. Suryadevi, D. Vijayakumar, R. Sabari Muthu Kumar, Dr. K .G. Srinivasagan , An Efficient Distributed Dynamic Load Balancing Algorithm for Private Cloud, International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 3, March 2014.

[2] Kyushu Zalavadiya, Dinesh Vaghela, Honey Bee Behavior Load Balancing of Tasks in Cloud, International Journal of Computer Applications April 2016.

[3] Kulkarni A.K., Annappa B. (2015). Load Balancing Strategy for Optimal Peak Hour Performance in Cloud Data centres., Signal Processing, Informatics, Communication and Energy Systems: SPICES 2015 IEEE International Conference, pp.1-5,Kozhikode ,ISBN NO.978-1-4799-1823-2/15, DOI: 10.1109/SPICES.2015.7091496

[4] James J., Verma B. (2012, September).Efficient VM Load Balancing Algorithm for a Cloud Computing Environment., International Journal on Computer Science and Engineering: IJCSE 2014, pp.1658-1663 ISSN:0975-3397,DOI: Sep 2012

[5] Access source - "https://www.javatpoint.com/history-of-cloud-computing", Accessed on 23 June 2019.

[6] Kamalakar.M, Moulika.T, A Priority Based Job Scheduling Algorithm in Cloud Computing, International Journal of Innovative Technologies Volume.03, IssueNo.01, May-2015.

[7] Mahalle H.M., Kaveri P.R., Chavan V. (2013, January) .Load balancing On Cloud Data Centres. International Journal of Advanced Research in Computer Science and Software Engineering: IJARCSSE 2013, India, pp.1-4, ISSN:2277 128X, DOI: Jan 2013

[8] Domanal S. G., Reddy G. R. M. (2013, October).Load balancing in Cloud Computing using Modified Throttled Algorithm. In Cloud Computing in Emerging Markets. Cloud Computing in Emerging Markets: CCEM 2013, pp.1-5, Bangalore, India, ISBN: 978-1-4799-0027-5, DOI: 10.1109/CCEM.2013.6684434

[9] Sharma T., Banga V.K. (2013, March).Efficient and Enhanced Algorithm in Cloud Computing. International Journal of Soft Computing and Engineering: IJSCE 2013 pp.385-390, ISSN: 2231-2307, DOI: 2013 March.