Cloud computing provides services to users to access SaaS, PaaS and IaaS services over the internet. There are many issues in Cloud computing like security, energy efficiency, big data, load balancing etc. One of the challenging issues is load balancing. Load balancing means distribute the workload to nodes or computers or resources so that we can achieve resource utilization, maximize throughput, reduce response time etc. Load balancing and job scheduling in cloud computing is very necessary for efficiency of resource utilization. Some algorithms are used for load balancing but some issues occur. So we can use another load balancing strategies for better load balancing and checks its efficiency through simulation results using simulator.

**Keywords:** Cloud Computing, Load Balancing, Job Scheduling, Resource Utilization.

---

### I. INTRODUCTION

Cloud refers to a different IT environment that is planned for the purpose of remotely provisioning scalable and calculated IT resources [1]. It is a kind of computing in which resource are shared rather than owning personal devices or local personal servers which can be used to handle applications on system. The word cloud in cloud computing is used as a symbol for internet so we can define a cloud computing as the internet based computing in which the different services like storage, servers and application are provided to organization computers and device using internet[2]. So as compare to conventional “own and use” method if we use cloud computing, the purchasing and maintenance cost of infrastructure is eliminated. It allows the users to use resources according to the coming of their needs in real time. Thus, we can say that cloud computing enables the user to have suitable and on-demand access of shared pool of computing resource such as storage, network, application and services etc base on pay per use.

#### A. Cloud Computing Architecture:

Cloud computing is growing in the real time environment and the information about the cloud and the services it provide and its deployment models are discussed. Figure 1 illustrating the three basic service layers that constitute the cloud computing. It provides three basic services that are Software as a Service, Platform as a Service and Infrastructure as a Service [2]. The rest of the paper is organized as follows. In section 2 we discussed virtualization of cloud. In section 3, load balancing and its necessity is discussed. In section 4 various load balancing techniques are discussed. And in section 5 Challenges of load balancing in cloud computing are explained.

![Fig. 1: Cloud Computing Architecture [1]](image-url)
II. CLOUD VIRTUALIZATION

In context of cloud computing the virtualization is very valuable concept. Virtualization is like “something that is not real” but provides all the services that are of real world. This is a software implementation of computer on which different programs can be executed as in the real machine. Virtualization is a component of cloud computing, because different services of cloud can be used by user. All these different services are provided to end user by remote data centers with full virtualization or partial virtualization manner [4]. There are two types of virtualization which are existing and it is describe under.

A. FULL VIRTUALIZATION

In full virtualization the complete installation of one system is done on other system. Due to this all the software that are present in real server will also existing in virtual system and also sharing of computer system among various users and emulating hardware placed on different systems are possible.

B. PARA VIRTUALIZATION

In this type of virtualization, many operating systems are allowed to run on a particular system by using system resources like memory and the processor (VMware software). Here total services are not fully available, but partial services are provided. Disaster recovery, migration and capacity management are some main features of Para virtualization.

III. LOAD BALANCING

Load balancing is one of the main issues related to cloud computing. The load can be a memory, CPU capacity, network or delay load. It is always necessary to share work load among the different nodes of the distributed system to improve the resource utilization and for improved performance of the system. This can help to avoid the condition where nodes are either heavily loaded or under loaded in the network. Load balancing is the process of ensuring the equally sharing of work load on the group of system node or processor so that without disturbing, the running task is finished. The goals of load balancing [7] are to:-
  - Improve the performance
  - Maintain system stability
  - make fault tolerance system
  - Accommodate future modification

There are mainly two types of load balancing algorithms:

A. STATIC ALGORITHM

In static algorithm the traffic is divided equally between the servers. This algorithm requires a earlier knowledge of system resources, so that the decision of shifting of the load does not depend on the present state of system. Static algorithm is correct in the system which has low variation in load.

![Load Balancing In Cloud Computing](image)

**Fig. 2 : Load Balancing In Cloud Computing [3]**

B. DYNAMIC ALGORITHM

In dynamic algorithm the lightest server in the whole network or system is look for and chosen for balancing a load. For this real time communication with network is required which can enhance the traffic in the system. Here present state of the system is make use of to build decisions to handle the load.
C. GOALS OF LOAD BALANCING
The load balancing of an application has a direct contact on the speedup. The rearrangement of balanced work-load by means of jobs and minimizing the inter process communication require with optimal resource utilization and task response time are the primary optimization objective of load balancing. Some of the main goals of a load balancing algorithm, as pointed out by [3] are to:

1) Performance Improvement:
Achieve a greater overall improvement in system performance at a reasonable cost, e.g., reduce task response time while keeping acceptable delays;

2) Job Equality:
To treat all jobs in the system equally regardless of their origin;

3) Fault Tolerance:
To have performance endurance under partial failure in the system;

4) Modifiability:
Have the ability to modify itself in accordance with any changes or expand in the distributed system configuration;

5) System stability:
The ability to account for emergency situations such as sudden surge of arrivals so that system performance does not deteriorate beyond a certain threshold while preventing nodes of the distributed system from spending too much time passing up jobs among themselves instead of executing these jobs.

IV. VARIOUS LOAD BALANCING ALGORITHMS

A. Round Robin Algorithm:
It would use of time slice method, as the name mean it works in a round approach where every node in the cloud environment is fixed with time slice and every node has to wait for their turn to execute their job. In other words it make use of random sampling method which signify the main controller choose the balancer arbitrarily to assign the load in case of some balancer is heavily loaded. When compared to other algorithm the complexity of round robin algorithm is less.[8]

B. Equally Spread Current Execution Algorithm
Equally spread current execution algorithm which allocates a job to every node with priority. It makes use of spread spectrum technique in which it share out the load over various nodes by read-through its load size. If the node is lightly loaded, the load balancer moves that job to that respective node and achieve high throughput. The load balancer keep queue of job selected to light weight node, which helps to recognize which node is free and need to be allocate with a new job.[8]

C. Throttled Load balancing
The Throttled algorithm will discover the specific node for assigning the new job. The job manager will keep a list of node detail using index list; with that it assign a particular job to particular node. If the node is ready to accept the particular job means it will accept and process otherwise it will wait for the other node requesting for processing. [8]

D. Connection Mechanism
Load balancing algorithm [9] can also be based on smallest amount connection mechanism which is a part of dynamic scheduling algorithm. It requires counting the number of connections for each server dynamically to approximation the load. The load balancer records the connection number of every server. The number of connection increases when a new connection is transmits to it, and decreases the number when connection finishes or timeout happens.

E. Min-Min Algorithm
It starts with a set of all unassigned tasks. First of all, least completion time for all tasks is found. Then among these least times the minimum value is selected which are the least time among the entire the tasks on any resources. Then according to that least time, the task is schedule on the related machine. Then the execution time for all other jobs is updated on that machine by adding the execution time of the assigned job to the execution times of other jobs on that machine and assigned task is eliminate from the list of the jobs that are to be assigned to the machines. Then once more the same process is followed until all the jobs are assigned on the resources. But this approach has a major disadvantage that it can lead to starvation [10].
F. Max-Min Algorithm
Max-Min is approximately same as the min-min algorithm apart from the following: after finding out minimum execution times, the maximum value is chosen which is the maximum time along with all the jobs on any resources. After that according to that maximum time, the job is scheduled on the related machine. Then the execution time for every other jobs is updated on that machine by adding the execution time of the assigned job to the execution times of other jobs on that machine and assigned job is removed from the list of the tasks that are to be assigned to the machines[10].

G. Biased Random Sampling
M. Randles et al. [11] proposed this distributed load balancing algorithm. Load balancing can be accomplished efficiently across the nodes in this approach, by using random sampling method. Here servers are worked as nodes. In this method a virtual graph is developed representing the load on the nodes and with every in-degree directed to the particular resources to the server. While the server starts executing the job it decreases the in-degree which specifies the decrease in availability of free resources. Likewise while the server completes the job the incoming degree gets increment which in turn indicates the raise in availability of resources. This process is called random sampling. The execution begins at any node and the random neighbouring node will be choosing for the next job to be executed. Thus the load balancing technique used here is fully decentralized and select apt for many cloud networks.[12]

H. Honey Bee Foraging
This algorithm was proposed by Dhinesh B.L.D, P.V.Krishna. This algorithm was originated from the behaviour of honey bees in finding their food [13]. Among the classes of bees the hunter bees hunt for food sources. Once the food source has been found the hunter bees come back to the bee hive and announce the food source by a dance called “waggle dance”. The kind of dance demonstrates the quality and quantity of the food and the distance of the source from the bee hive. This makes the survey bees to race for the food. In case of load balancing the servers are grouped into virtual servers. Each Virtual servers will has its own Virtual Server (VS) request queue. A server provide a request, compute its profit and evaluate it with the colony profit, if profit was high, then the server live at the existing virtual server and on the other hand if profit was low, then the server returns to the hunt or survey behaviour, thus balancing the load with the server.

I. Active Clustering
Active clustering is an improved method of random sampling, where this algorithm works on the principle of grouping related nodes together and start working on these group nodes [10]. This method uses the resources efficiently so enhance the throughput and performance of the system by using high resources. In this approach a method called match-maker is introduced. When an execution starts in a network, the process gets initiates and finds for the next similar node said to be match-maker which should satisfy the criteria that it should be the dissimilar one from the former one. Once the match-maker is found the process gets initiated and as soon as the process gets over the match-maker gets separate from the network. Thus this is an iterative process in the network to balance the load efficiently.

V. CONCLUSION AND FUTURE WORK
In this paper, we have surveyed various load balancing techniques for cloud computing. The main purpose of load balancing is to satisfy the customer requirement by distributing load dynamically among the nodes and to make maximum resource utilization by Reassigning the total load to individual node. This ensures that every resource is distributed efficiently and evenly. So the performance of the system is increased. We have also discussed virtualization of cloud and required qualitative matrix for load balancing.

ACKNOWLEDGMENT
The authors would like to thank the reviewers for their precious comments and suggestions that contributed to the expansion of this work.

REFERENCES
A Survey on Load Balancing and Scheduling in Cloud Computing


