LOAD BALANCING

A Classic Technique For Resource Management

[1]Bhoomi Gupta, [2]Nikita Gupta, [3]Mayank Gupta

[1]Assistant Professor, [2]Student, [3]Student
[1]Department of Information Technology,
[1]Maharaja Agrasen Institute of Technology, Delhi, India

Abstract: **This paper aims at cloud computing - an inevitable asset and one of the emerging areas in the field of information technology (IT)- and its most important component, load balancing. Cloud gained popularity because of its ability to access resources from anywhere and pay as you go model. The increasing applications of cloud in web applications and systems have led to major developments in the field of load balancing and management in the past few decades. Scaling web applications is a major issue that is faced by many companies and cloud provides an effective way of resolving this. Load balancing is a classic technique used to balance out the workload among various nodes for better utility and resource management, thus increasing the efficiency of the system.**

Key Words: Load Balancing, Cloud Computing, Ant Colony Optimization,

1. Introduction to Cloud Computing

1.1 Definition of Cloud Computing

With the success of the Internet, rapid development was seen in the field of processing and storage technologies, which resulted in computing resources becoming cheaper, more powerful and more omnipresent than ever before. This technological trend has enabled the realization of a new computing model called cloud computing, in which resources (e.g., CPU and storage) are provided as general utilities that can be leased and released by users through the Internet in an on-demand fashion. (Zhang, Cheng, & Boutaba, 2010) [1]

The term “cloud” gained popularity in 2006 after it was used by Eric Schmidt, to describe the business model of providing services across the Internet. The lack of a standard (conclusive) definition of cloud computing has resulted not only in market hype but also a fair amount of skepticism and confusion. The work in (Vaquero, Rodero-Merino, Caceres, & Lindner, 2009) [2] focused on comparing 20 different definitions of cloud computing from various sources to assert a standard definition. However, in this paper, we accept the definition provided by The National Institute of Standards and Technology (NIST), “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.” [3]

Cloud computing services provided can be understood as a one-stop solution for all the requirements for any organization. Few distinct computing services required are:

- Compute power – to scale and innovate
- Storage services – database storage and management
- Networking services – handle the connection between physical and private networks
- Analytics- to analyze data using techniques like Machine Learning.

1.2 Service and Deployment Model

The four types of deployments models that have been defined for the cloud community are as follows:

1.2.1 Private Cloud: The entire infrastructure of the cloud is employed for a single business or organization, it may be located on or off the organization premises and can be operated by either the organization or any third party. (Dillon, Wu, & Chang, 2010) [4]

1.2.2 Public Cloud: It is one of the most prepotent forms of Cloud computing deployment model. It is used by the general public cloud consumers. The public cloud is completely owned and managed by the cloud service provider with its own set of policies, values, costing, and charging model. Amazon EC2, Google AppEngine and Force.com, are amongst some of the most popular public cloud services in the market today. (Dillion et al., 2010)[4]

1.2.3 Community Cloud: The cloud infrastructure is jointly constructed by several organizations together and can be hosted by a third party cloud provider or within the premises of one of the organizations. The community shares all the concerns, values, requirements and policies of the cloud. (Dillion et al., 2010) [4]

1.2.4 Hybrid Cloud: This type of cloud infrastructure is the integration of two or more cloud computing service models (private, public, or community). (Dillion et al., 2010) [4] It tries to overcome the limitations of each approach. It follows the approach where a part of the cloud runs on a public cloud while the other runs of private cloud. (Zhang et al., 2010)[1] Businesses use hybrid cloud for resource optimization and hence manage their data efficiently.

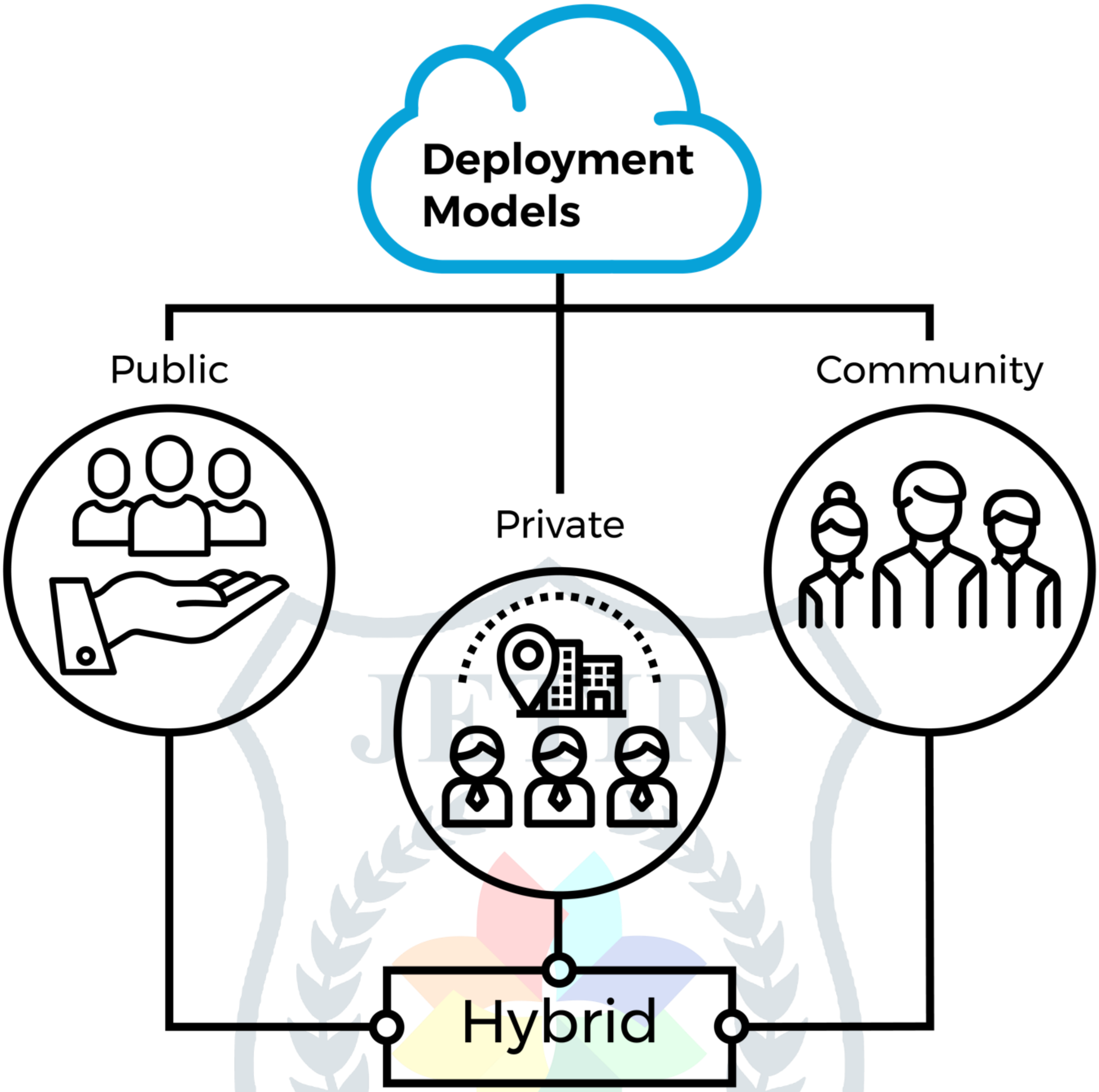

Figure 1: Cloud Deployment Models

Apart from these 4 deployments models, the cloud can also be classified into 3 service models:

1.2.5 *Software as a Service* (SaaS): It is the on-demand provisioning of applications over the internet. These applications can be accessed through various client devices, usually using a web browser. (Zissis, & Lekkas, 2012) [5] Example: Google Apps such as Mail, Docs, and Spreadsheets (Patidar, Rane, & Jain, 2012)[6]
1.2.6 *Platform as a Service* (PaaS): It is the provisioning of platform services, like network, servers or operating systems, and is aimed at developers. The upscaling of the application is managed by the cloud service provider as the usage of the application grows. (Patidar et al., 2012)[6] Example, Google AppEngine and Force.com
1.2.7 *Infrastructure as a Service* (IaaS): It is the on-demand provisioning of processing, storage, and other fundamental infrastructural resources usually as VM's (Virtual Machines). (Zhang et al., 2010) [1] These are provided to the consumer by a cloud service provider. Example, Amazon Web Services with its Elastic Compute Cloud (EC2). (Dillion et al., 2010) [4]

2. Cloud computing challenges

2.1 Security: Security and privacy are the biggest all-time concerns of the cloud computing community. Some companies hesitate to hand over their confidential data and adopt cloud services. The ability to access this data from anywhere makes the possibility of clients privacy being compromised quite high. One way to overcome this issue is to employ proper authentication techniques.
2.2 Costing Model: However migrating to the Cloud can help significantly reduce the infrastructural cost, but this, in turn, raises the data communication cost, i.e. the cost required for transferring the company's data to and from the public and community cloud. (Youssef, & Ahmed, 2012) [7] The cost of every computing resource such as a VM is also expected to be higher. (Dillion et al., 2010) [4]

2.3 Charging Model: From a cloud provider's perspective, the elastic resource pool (through either virtualization or multi-tenancy) has made the cost analysis a lot more complicated than regular data centers, which often calculates their cost based on consumptions of static computing. Moreover, an instantiated virtual machine has become the unit of cost analysis rather than the underlying physical server. A sound charging model needs to incorporate all the above as well as VM associated items such as software licenses, virtual network usage, node and hypervisor management overhead, and so on. (Dillion et al., 2010) [4]

2.4 Scalability Issues: Scalability in simple terms means the ability of any system to perform well under an increased or dynamic workload. A system that scales well will be able to maintain or even increase its level of performance or efficiency when tested by larger operational demands (Tripathi, & Singh, 2017) [9].

There are two types of scaling:

a. Scale Up: the ability to handle increased workload efficiently.

b. Scale Down: the ability to manage tasks that require fewer resources

Thus, scaling up and down requires a flexible mechanism to enable this functionality to optimize resource usage and reduction in costs.

3. Introduction to load balancing

Cloud computing is one of the most emerging and prominently used internet-based technology that emphasizes commercial computing. (Mishra, & Mishra, 2015) [10] It is a paradigm that's aimed at providing businesses with dynamic and scalable virtual computing resources over the Internet on demand and uses pay as go model. It is a further extension of parallel, distributed and grid computing. (Fang, Wang, & Ge, 2010)[11] Overloading and underloading of the cloud service is a major issue faces by the cloud services providers(CSP's), as it may lead to failures along with poor power consumption, increased execution time, machine failures, etc. Thus to cater to this problem a new approach called load balancing was introduced. Load balancing is a relatively new technique that facilitates networks and resources by providing a maximum throughput with minimum response time (Zenon, Venkatesh, Shahrzad and Christopher, 2011) [12]. The aim behind load balancing is the distribution of load (client requests sent to the server) among the nodes of the cloud infrastructure evenly. This would lead to better response time, execution time along with the reduction in costs (Deepa, & Cheelu, 2017) [13].

3.1 Definition of load balancing

Load balancing is a mechanism that distributes the workload (user requests) evenly across all the nodes in the entire cloud to avoid any situations, where some nodes are heavily loaded (overloaded) while the others are idle or doing less work(underloaded). The workload of a machine means the total processing time it requires to execute all the tasks assigned to the machine. (Sreenivas, Prathap, & Kemal, 2014) [15] It helps by providing higher user satisfaction and better resource utilization ratio, hence improving the overall performance and resource utilization. It averts any gridlocks of the system that might occur due to load imbalance. Load balancing helps in the continuation of the normal system functioning, even if one or more components fail by applying fail-over, i.e switching to standby or ideal computer server, network or hardware upon failure or abnormal termination. (Kansal, & Chana, 2012) [14]

Some of the major goals of load balancing algorithms to balance the requests from the resources are:

a. **Cost-effectiveness:** The overall system performance should improve at a reasonable cost

b. **Scalability and flexibility:** There might be some size and topology changes in the distributed system in which the algorithm is implemented. So the algorithm must be dynamic in order to allow these changes easily.

c. **Priority:** Resource and job prioritization must be done by the algorithm itself so that better service is offered to important or high priority jobs irrespective of equal service provision and origin of a job (Ray, & Sarkar, 2012) [16].

3.2 Advantages of Load Balancing

Following are the advantages of load balancing (Tripathi et al., 2017)[9]:

a) Better performance of applications

When comparing traditional systems with on-site servers and cloud load balanced systems, we find that the latter is cheaper with easy implementation. Thus, helping an organization using load balancing techniques by delivering applications that are faster and are more efficient at lower costs.

b) Boosting scalability

Cloud services provide easy scalability and celerity of website traffic management. Now, load balancing boosts these abilities, as now it is easier to distribute the traffic among various unallocated nodes. This is very crucial when it comes to websites like that built for e-commerce dealing with a lot of viewers every second. Especially during sales or other promotions, effective balancing of the load is required to distribute the workload.

c) Handling increased traffic

Consider a simple example like a website declaring results for any national or university-level examination. A website simply working on limited resources will go down as it will not be able to handle the sudden increase in traffic. Load balancers help resolve this issue by redistributing the requests to free nodes. Thus, making the process faster and smoother.

d) Deals with sudden outages and manages functions

A load balancer protects the website from sudden glitches. This can be explained by the distribution of workload among the many servers so that when one node fails, there are substitute nodes available. Thus, load balancers provide an all-round solution with scalability, traffic control, and outage management.

4. Load Balancing metrics

Load balancing is the reason behind the smooth functioning of cloud computing networks by distributing the load among the nodes available in the network fairly. The choice of algorithm for load balancing depends upon a variety of factors mentioned below (Tripathi et al., 2017) [9]. We require higher throughput, response time and performance. We also require scalability, efficient resource utilization and fault tolerance.

1. Throughput

It basically refers to the number of tasks that have been successfully completed. It is required to be high. (Nishat et al., 2012) [17].

2. Response Time

This parameter refers to the amount of time taken by any load balancing algorithm to respond to a request made by a client in the network. It should be as low as possible to make the system efficient (Nishat et al., 2012) [17].

3. Scalability

Scalability refers to the ability of the network to be able to handle a finite number of nodes. This should be such that the system becomes flexible to maximize efficiency and minimize the computation involved (Nishat et al., 2012) [17].

4. Resource Utilization

It refers to the optimal utilization of the resources in the network. It should be as high as possible (Tripathi et al., 2017) [9].

5. Overhead

Overhead refers to the communication overhead that results from the movement of tasks, communication between processors and any other communication in the network. The lesser the overhead, the more efficient the performance of the algorithm (Nishat et al., 2012) [17].

6. Performance

It is a more holistic parameter which depends on the factors mentioned above. A system with good performance is one that has high throughput, scalability, resource utilization, low overhead and low response time.

5. Load Balancing Algorithms

As the cloud computing technology is gaining popularity in the expansion of businesses, load balancing becomes a compulsory mechanism to employ for fair distribution of workload among the nodes in the network. There are numerous algorithms employed for load balancing, and to choose the most appropriate algorithm for our network environment. The choice of the algorithm requires a complete understanding of the various algorithms and their benefits and drawbacks. Let us have a look at the classification hierarchy of load balancing techniques.

5.1 Classification of algorithms – based on nature and distribution

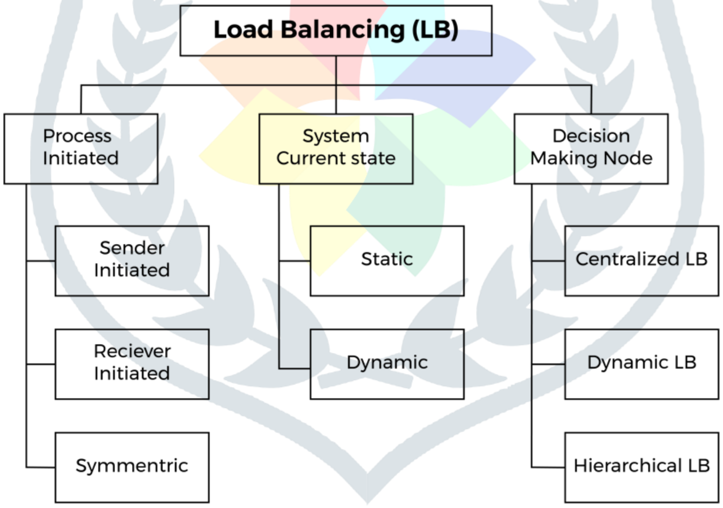

Figure 2: Load Balancing Classification

1. Based on Process Initiation

Depending upon who initiated the process of load balancing, the load balancing strategies can be divided into the following categories:

a. Sender-Initiated

The sender-initiated algorithms are those where the client initiates the request and then a receiving node caters to the request of load sharing (Mishra et al., 2015) [10].

b. Receiver-Initiated

The receiver node sends the acknowledged request to the sender that will then initiate the load balancing process (Mishra et al., 2015) [10].

c. Symmetric

This is a combination of sender-initiated and receiver-initiated types.

2. Based on System Current State

Based on the current state of the network, the load balancing strategies can be divided broadly into two categories:

a. Static algorithms

Static algorithms are those algorithms where the knowledge base consists of the previous knowledge of the statistics of each node and the client needs (Tripathi et al., 2017) [9]. We should have thorough knowledge about the existing system like memory, performance, and processing. They do not consider the current system state and work according to an already set model. They are generally employed in a homogeneous environment (Tripathi et al., 2017) [9].

Advantages:

The main advantage of using this strategy is that the computation time is less as the current state is not maintained and needed in the process.

Disadvantages:

In the case of system failures and attempts to shift tasks during the execution of static algorithms leads to serious issues. This is due to the fact that they work in a non-pre-emptive manner (Mishra et al., 2015) (Mishra et al., 2015) [10]. Scalability and flexibility issues are the main limitations to using this strategy. An example of this strategy is the Round Robin, max-min and opportunistic load balancing techniques.

b. Dynamic algorithms

These are those algorithms that incorporate the run-time statistics of every node to calculate the constant load-changing requirements. They react to the changing system state dynamically.

Advantages:

They avoid degradation of performance by preventing the computation of unnecessary nodes by decreasing the states. They also provide better fault tolerance (Mishra et al., 2015) [10].

Disadvantages:

More complex as compared to static algorithms as they keep track of the states and the computation time required to process them is higher.

They are widely used in distributed systems and some examples are ant colony optimization and honey bee foraging.

3. Based on the Decision- Making Node

Depending upon which node makes the decisions regarding load balancing, there are three types of load balancing strategies.

a. Centralized Load Balancing

In this type of load balancing algorithms, a single node performs the task of allocating load and performing schedule. This node is called the central node. The entire network is dependent on this particular node and it holds the knowledge base for both- static as well as dynamic load balancing. It makes it faster to compute and allocate resources but the fact that the entire system is dependent on a single node makes the network vulnerable to failures. Thus, resulting in poor fault tolerance and difficult recovery from system failures (Mishra et al., 2015) [10].

b. Distributed Load Balancing

In contrast to the centralized load balancing technique, here the decisions regarding task scheduling and load distribution are not made by a single node, but a collection of domains. A knowledge base is created in a manner similar to that of static load balancing but the redistribution is done dynamically. Thus, more efficient and more fault tolerance is achieved (Mishra et al., 2015) [10].

c. Hierarchical Load Balancing

Slave mode operation technique is used in the hierarchical load balancing technique. The load balancing process is carried out at various levels. This concept can be well implemented using a data structure like a tree where every child node has a parent node. The parent node makes the decisions for the child node (Mishra et al., 2015) [10].

6. Load Balancing Techniques

6.1 Ant Colony Optimization

Ant Colony Optimization (ACO) is a widely used algorithm which is based on genetic algorithms and considers the natural behaviour of ants to allocate the load among nodes. The ant colonies working together and their foraging nature have always inspired researchers to use this in their experimentations to solve problems. The ants work together as a network while searching for food and use the existing paths to transfer food back to the nest. While looking for the methodology used by blind ants to follow paths, researchers found a substance called pheromone that is used by ants for navigation (Kansal et al., 2012) [14]. The pheromone trails led ants to the right path. These pheromone trails are used by ants to follow each other. They even update the trails to make it easier for traversal. The paths with maximum pheromone trails are the shortest ones between the point and food source. The same idea is used in load balancing also (Kansal et al., 2012) [14].

The movements while traversal is categorized into:

- Forward movements – these are the movements made by ants to search or collect food from sources.
- Backward movements – these movements are used to traverse back to the nest after collecting the food from sources.

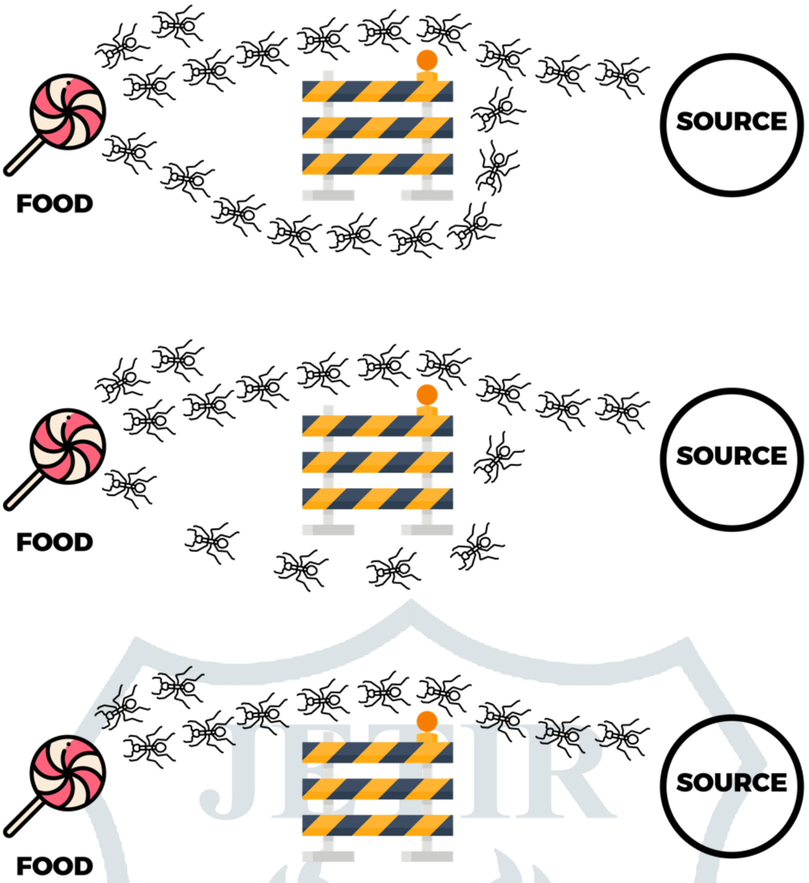

Figure 3: Ant colony optimization

6.2 Honeybee Foraging Algorithm

Honeybee Foraging is another decentralized load balancing method inspired by nature where the network design is based on the behaviour similar to that of bees and assist in balancing the load across the nodes. (Kaur, & Luthra, 2014) [18] The first step is searching for a candidate node i.e. an overloaded or underloaded node. The next and most crucial step of the algorithm is to distribute the load from an overloaded node to an underloaded node considering its priority.

Honeybee foraging

Bees live in colonies and search for their food – nectar or pollen from flower patches. Some of the bees are classified as the worker bees who look for food sources and others perform the task of following the worker bees. When food sources are found, the scout bees go to the field surrounding the hive and collect the harvested food. There is a special waggle dance performed by the scout bees once they find a beneficial food source. This dance is responsible for communicating the source to an idle bee. The duration of dance is directly proportional to the scout's rating of the source of food found. Thus, once this good food source needs to be harvested, more foragers are recruited (Gupta, & Sahu, 2014) [20] (Zenon et al.) [12].

This algorithm is commonly used when a population comprising of varied service types is needed. A drawback in employing this algorithm is that as the system size increases, there is no corresponding rise in the throughput (Jadhav, 2012) [19].

6.3 Active Clustering

Active clustering is a load balancing algorithm that includes the concept of self-aggregation which involves the grouping together of the similar type of tasks and working them together in groups. The optimization of assignments of similar jobs occurs by connecting similar services (Kaur et al. 2014) [18]. The main advantage of using this algorithm is increased throughput and system efficiency. It uses resources optimally providing a stable environment to work on.

This algorithm is used in cloud computing environments that require working with groups of similar tasks. The main aim of the assignment of groups is to optimize task assignment and decrease resource usage for multiple assignments. It is a better version of the random sampling theorem as it uses the concept of clustering.

Matchmaker node – A node that begins the process of clustering and chooses another process is called a matchmaker node.

6.4 Round Robin

The method of distribution works by moving down the servers and by forwarding a request to each server according to the sequence (Etminani,& Naghibzadeh, 2007) [21]. After reaching the last server in the list, it again starts with the first and through the entire list again.

The main advantage of the Round Robin algorithm is its simplicity of implementation (Etminani et al., 2007) [21]. It might not always result in the correct traffic distribution. This is due to the reason that it assumes that all the servers are the same. There are two variants of this algorithm:

- **Weighted round robin** – In the weighted round robin, each server in the network is assigned a weight signifying the proportion of client requests each participating server receives. For example, if a server A is allotted a weight of 2 and a server B, a weight of 1, 2 requests will be forwarded to server A and 1 to server B.

- **Dynamic round robin** – Dynamic assignment of weights to each server based on the existing load and the actual capacity of the server.

References

[1] Q. C. L. &. B. R. Zhang, "Cloud computing: state-of-the-art and research challenges.," *Journal of Internet Services and Applications, 1(1), 7–18.* , 2010.

[2] R.-M. L. C. J. L. M. Vaquero L, "A break in the clouds: towards a cloud definition," *ACM SIG- COMM computer communications review* , 2009.

[3] "“The NIST Definition of Cloud Computing,” National Institute Of Standards and Technology.".

[4] T. W. C. &. C. E. Dillon, "Cloud Computing: Issues and Challenges," *IEEE International Conference on Advanced Information Networking and Applications,* 2010.

[5] D. &. L. D. Zissis, "Addressing cloud computing security issues," *Future Generation Computer Systems,* 2012.

[6] S. R. D. &. J. P. Patidar, "A Survey Paper on Cloud Computing," *International Conference on Advanced Computing & Communication Technologies.,* 2012.

[7] A. E. Youssef, "Exploring Cloud Computing Services and Applications," *Journal of Emerging Trends in Computing and Information Sciences,* 2012.

[8] Y. &. M. K. Jadeja, "Cloud computing - concepts, architecture and challenges," *International Conference on Computing, Electronics and Electrical Technologies (ICCEET),* 2012.

[9] S. S. Aanjey Mani Tripathi*, "A literature review on algorithms for the load balancing in cloud computing environments and their future trends," *COMPUTER MODELLING & NEW TECHNOLOGIES,* 2017.

[10] N. M. Nitin Kumar Mishra, "Load Balancing Techniques: Need, Objectives and Major Challenges in Cloud Computing- A Systematic Review," *International Journal of Computer Applications,* vol. 131, 2015.

[11] Y. W. F. &. G. J. Fang, "A Task Scheduling Algorithm Based on Load Balancing in Cloud Computing," *Lecture Notes in Computer Science, 271–277. (springer)* , 2010.

[12] V. M. S. A. a. C. M. Zenon Chaczko, "Availability and Load Balancing in Cloud Computing," *2011 International Conference on Computer and Software Modeling IPCSIT,* vol. 14, 2011.

[13] D. D. C. T. Deepa, "A Comparative Study of Static and Dynamic Load Balancing Algorithms in Cloud," *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS),* 2017.

[14] I. C. Nidhi Jain Kansal, "Cloud Load Balancing Techniques : A Step Towards Green Computing," *IJCSI International Journal of Computer Science Issues,* vol. 9, 2012.

[15] V. P. M. &. K. M. Sreenivas, "Load balancing techniques: Major challenge in Cloud Computing - a systematic review," *International Conference on Electronics and Communication Systems (ICECS),* 2014.

[16] A. D. S. Soumya Ray, "EXECUTION ANALYSIS OF LOAD BALANCING ALGORITHMS IN CLOUD COMPUTING ENVIRONMENT," *International Journal on Cloud Computing: Services and Architecture (IJCCSA),* vol. 2, 2012.

[17] P. S. V. K. C. G. a. K. P. S. N. a. R. R. K. Nishant, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization," *International Conference on Modelling and Simulation,* 2012.

[18] R. Kaur, "Load Balancing in Cloud Computing," *Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC.*

[19] S. P. Jadhav, "Load Balancing in Cloud Computing," *International Journal of Science and Research (IJSR), no. ISSN (Online): 2319-7064* , 2012.

[20] K. S. H. Gupta, "Honey Bee Behavior Based Load Balancing of Tasks in Cloud Computing," *International Journal of Science and Research (IJSR) , Vols. ISSN (Online): 2319-7064.*

[21] K. E. a. M. Naghibzadeh, "A Min-Min Max-Min Selective Algorihtm for Grid Task Scheduling," *3rd IEEE/IFIP International Conference in Central Asia,* 2007.