A COMPARATIVE SURVEY ON EXISTING LOAD BALANCING ALGORITHMS IN CLOUD

Prof. Kaushal Gor
Assistant Professor
Department of MCA
Parul University

Prof. Vibhuti Patel
Assistant Professor
BCA Department
Parul University.

Abstract: Cloud is a popular & efficient platform which provides various on-demand services such as computing & storage. It provides resources, shared information, software packages & other resources as per user's needs. To enhance the performance & utilization of resources in cloud, load balancing is required. The purpose of this survey paper is to provide brief information about various load balancing algorithms of cloud. This survey helps to analyze various algorithms on the basis of different qualitative metrics.

Index Terms **– Cloud Computing, Load, Load Balancing, Load Balancing Algorithms.**

INTRODUCTION

Cloud Computing

It is a new paradigm where computing & storage resources are made available to the users as pay per use model. It is a distributed computing model that provides its users with a shared resource, information, virtualized hardware and/or software infrastructure over the internet. It has an ability to run an application on many connected systems at the same time via internet. It offers online data storage, compute infrastructure and application on demand. Example: Google.

Cloud service provider may have offered these services as free of cost e.g. Google. On other hand services may offered services as pay per use model [1]. Example: Microsoft Azure, Amazon AWS.

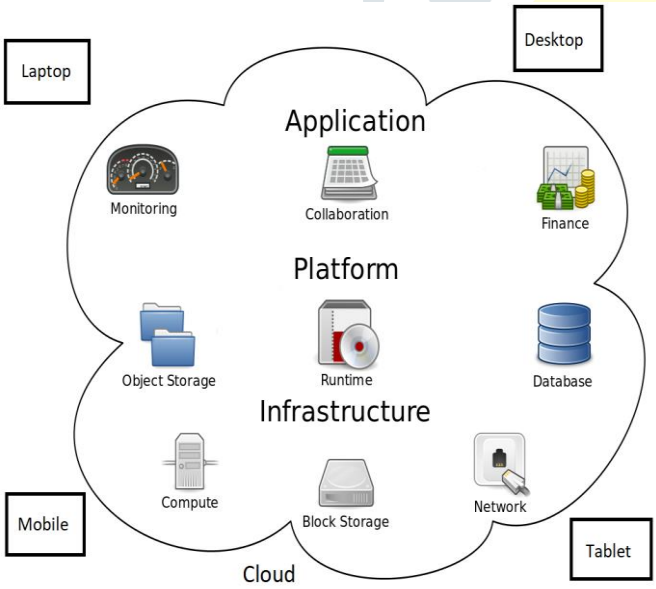

Fig. 1 Cloud Environment

There are mainly four types of cloud deployment model are available namely public cloud model, private cloud model, community cloud model & hybrid cloud model. In public cloud environment, resources are accessible to public. Example: Amazon AWS, Google. In private cloud, it allows systems & services to be accessible within an organization. It is more-costly than public cloud. Example: VMWare & Rackspace. In community cloud, it allows systems & services to be accessible to several organizations with a common approach. Example: Salesforce. And hybrid cloud model is a combination of community, private & public cloud where sensitive activities are performed using private cloud & other activities are performed using the public cloud. Examples: Egenera PAN Cloud Director.

Cloud Computing provides mainly three types services namely SaaS (Software as a Service), PaaS (Platform as a Service) & IaaS (Infrastructure as a Service). In SaaS, it offers software application as a service to the customer. Example: Google Apps and Microsoft Office 365. In PaaS, it provides a runtime environment for applications, development & deployment tools, etc. Application developers can develop & run the software on a cloud environment. Cloud Foundry and open shift origin are open source implementation for PaaS. Examples: Windows Azure and AWS Elastic Beanstalk. In IaaS, Client can access the resources over the network such as processing, networking, storage and other computing resources. Examples: Amazon EC2 and Windows Azure.

Load Balancing

Load balancing uniformly distribute the work load on virtual machines for proper utilization of resources. Its main goal is to provide continuous service in case of failure of any service component by provisioning the application instances along with proper utilization of resources.

Load balancing aims to minimize response time for tasks & provide better resource utilization, which increase the system performance at low cost.

Load balancing performs two major things: one is resource allocation and job scheduling which ensures the availability of resources, efficient resources utilization, low load, reduced cost of resources.

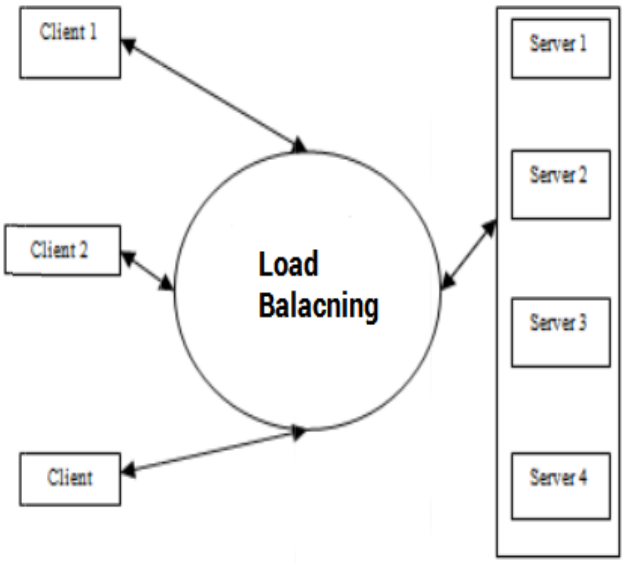

Fig. 2 Load Balancing [1]

For efficient load balancing, a load balancer is required. Load Balancer helps in fairly allocation of resources for resource utilization & user satisfaction with minimal cost. It also ensures a minimal response time which results in maximum throughput [1].

A load balancer is like a traffic distributer that distributes client requests across the clouds & check that no nodes are overloaded. The request is redirected to the idle node if a single node gets overloaded. The load balancer automatically starts working with a newly added node into the cloud [1].

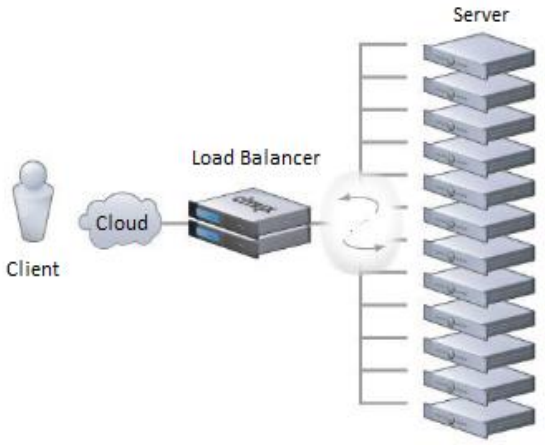

Fig. 3 Working of Load Balancer

Proper distribution of load among the nodes reduces network load, improve performance & services such as fast response time, user satisfaction, optimum cost and data backup & recovery.

There are various algorithms available for load balancing in cloud computing environment. Study & survey of various existing balancing algorithms is provided in next section. A comparative survey is also provided mentioning behavior of algorithm with reference to particular parameters.

SURVEY & WORK ANALYSIS

The load balancing algorithms can be defined into static & dynamic algorithms.

Static Algorithms

Static load balancing do not require knowledge of the current state of a system. It keep prior knowledge about resources such as memory, storage capacity & execution time. Hence, the judgment of load distribution does not rely on the present state of the system. Such algorithms face a foremost drawback in case of sudden failure of system resource and tasks [1][3]. So this algorithm is not useful for performing distributed computing.

Algorithms in Static Load Balancing

A. Round Robin (RR) Algorithm:

In RR algorithm fixed quantum time is given to request. RR emphasis on fairness & time limitation. In RR algorithm user requests are processed in a circular manner. It utilizes equal period to complete each task [5] [6]. In case of heavy load, round robin takes a long time to finish all the given tasks. In case of larger tasks, it takes longer time for completion [7].

B. MIN-MIN Algorithm:

In this algorithm, it requires prior knowledge of system resources & available jobs. This algorithm selects the job which has minimum completion time from the set of all pending job. After selecting a job, the node with minimum completion time is selected. Finally, Assigning the selected job to the selected node. The process repeats until all the pending jobs are completed. The advantage of MIN-MIN algorithm is the job having smallest completion time is executed first [8].

C. MAX-MIN Algorithm:

MAX-MIN algorithm considers the maximum completion time of nodes. The algorithm requires prior knowledge of system resources and available jobs. This algorithm selects the job which has minimum completion time from the set of all pending job. After selecting a job, the node with maximum completion time is selected. Finally, Assigning the selected job to the selected node. The process repeats until all the pending jobs are completed. [8]

D. Shortest Job Scheduling (SJS) Algorithm:

In SJS algorithm [10] shortest executable job is selected first. This algorithm follows to perform the complete execution of short jobs to utilize the resources in completion of heavy jobs.

Dynamic Algorithms

In previous section static algorithms do not need knowledge of the current state of the system that's why these techniques are not suitable for distributed computing systems.

The dynamic algorithm does not require any prior information about system resources because the load distribution decision is based on the current state of the system [1]. Dynamic load balancing algorithms distributes load at runtime. Dynamic load balancing algorithms are complex in nature but have better fault tolerance & high performance [3]. Such algorithms provide better performance improvement the static algorithms [4].

Different policies are used for the dynamic load balancing.

- Transfer Policy: Selecting job to transfer from local node to remote node.
- Selection Policy: This policy used for load exchange specified the processors.
- Location Policy: Location policy describes the destination node for transferring.
- Information Policy: Information policy is Collects the information of the node in the network.

Algorithms in Dynamic Load Balancing

A. Genetic Algorithm (GA):

GA is a part of evolutionary computing that evolves from generation to generation. GA is used in homogenous. GA is a search heuristic that matches the process of natural selection [11].

In GA, the individuals in a population compete for resources. The fittest individuals are able to acquire the required resource. The good individual which satisfies the survival condition is allowed to propagate to further generation. Hence, minimizes the time to give profit to cloud service provider and reduce resource maintenance cost.

GA focuses on following steps:

- **Population:** The number of nodes over the network is considered as population.
- **Selection Function:** Light weighted nodes according to the service requests.
- **Crossover:** Balancing a load of a heavy weighted node by transferring it to light weighted node.
- **Fitness Function:** Performance and Resource Utilization of the nodes.

GA uses centralized balancing approach that provides more efficiency than static algorithm.

B. Ant Colony Algorithm (ACO):

ACO is the heuristic approach for load balancing [11]. Introduced a scheduling strategy based on Load Balancing Ant Colony Optimization (LBACO) algorithm where they balance the system load while trying to reduce the makespan.

Dorigo *et al* [12] proposed an approach ACO which is based on nature of real ants, which form a network in order to process the job. ACO is inspired by the path exploring actions of various ants species. For searching the path, ants communicate with each other using a liquid the ants evaporates called pheromones. During this, other ants follow the same path with the help of released pheromones. The likelihood of ant chooses a way is directly proportional to the concentration of the pheromones in that path. The path having dense pheromone has more chances that the ants choose that path. With the help of intensity of pheromones, ants can find an optimal path.

Similar to the searching behavior of the real ants, in cloud computing ACO searches the light weighted nodes over the network and based on the load of the nodes it distributes the load in the network.

C. Active Clustering Algorithm:

Active Clustering is improved method of random sampling. The concept of clustering is used in this. The main principle of this algorithm is grouping similar nodes together & working based on those nodes. Grouping of nodes helps the resources to increase the throughput efficiently. In this algorithm, a method called matchmaker is introduced [13].

D. Throttled Algorithm:

In [14], authors have described algorithm in which the required operations are performed based on the request made by the client to load balancer for searching appropriate virtual machine. These VMs can be grouped based on the type of requests, as per the incoming requests it works accordingly. Each time a user sends a request, the load balancer will search for the related type of group and if such a group is ready to accept and handle the request, the request will be assigned.

E. Honey Bee Foraging Algorithm:

The Honey bee behavior inspired load balancing algorithm basically manages the load across different virtual machines for increasing throughput. Tasks are prioritized so that the waiting time is reduced when they are aligned in queues. The honey bee foraging behavior & some of its variants are listed in [16]

Based on the various criterions - Type of algorithm, Performance, Resource utilization, Response time, Throughput and Scalability - The above mention algorithms are classified in Table I [15].

Table I. Comparison of various Load Balancing Algorithms based on various Metrics

Load Balancing Algorithm	Type	Overhead	Performance	Resource Utilization	Response Time	Throughput	Scalability
Round Robin	Static	Yes	Yes	Yes	Yes	Yes	Yes
MIN-MIN	Static	Yes	Yes	Yes	Yes	Yes	No
MAX-MIN	Static	Yes	Yes	Yes	Yes	Yes	No
Shortest Job Scheduling	Static	Yes	No	Yes	No	No	No
Genetic Algorithm	Dynamic	Yes	Yes	Yes	No	No	No
Ant Colony	Dynamic	No	Yes	Yes	No	No	No
Active Clustering	Dynamic	Yes	No	Yes	No	No	No
Throttled	Dynamic	Yes	Yes	Yes	Yes	No	Yes
Honey Bee Foraging	Dynamic	No	No	Yes	No	No	No

CONCLUSION

There are many problems in cloud computing, one of them is Load Balancing. Through Load Balancing workload is distributed over the network. It reduces the dynamic workload across all nodes for the achievement of higher user satisfaction & resource allocation. The purpose of load balancing is to make every node perform the same amount of work throughout which helps in increasing throughput, reducing the response time and the number of job rejection. In this paper comparison of various Load Balancing algorithms are discussed, there are many other Load Balancing algorithms are available that can be used. A large number of parameters & different types of soft computing techniques can be included in the future for the better utilization.

REFERENCES

[1] Tushar Desai & J. Prajapati "A survey of various load balancing techniques and challenges in cloud computing." International Journal of Scientific & Technology Research 2.11 (2013): 158-161.

[2] Sidra Aslam & Munam Ali Shah, "Load Balancing Algorithms in Cloud Computing: A Survey of Modern Techniques" National Software Engineering Conference (NSEC 2015).

[3] N. S. Raghava & Deepti Singh. "Comparative study on load balancing techniques in cloud computing." Open Journal of Mobile Computing and Cloud Computing 1.1 (2014).

[4] Abhijit A. Rajguru & Mrs. Sulabha S. Apte. "Various Strategies of Load Balancing Techniques and Challenges in Distributed Systems."

[5] N. Sran & N. Kaur, "Comparative analysis of existing load balancing techniques in cloud computing" International Journal of Engineering Science Invention, vol. 2, no. 1, pp. 60–63, 2013.

[6] S. Subashini & V. Kavitha, "A survey on security issues in service delivery models of cloud computing" Journal of network and computer applications, vol. 34, no. 1, pp. 1–11, Jan. 2011.

[7] N. J. Kansal & I. Chana, "Existing load balancing techniques in cloud computing: a systematic review" Journal of Information Systems and Communication, vol. 3, no. 1, pp. 87–91, 2012.

[8] Nitin Kumar Mishra & Nishchol Mishra, "Load Balancing Techniques: Need, Objectives and Major Challenges in Cloud Computing- A Systematic Review" International Journal of Computer Applications (0975 – 8887), vol. 131.

[9] T. Kokilavani and Dr. D.I. George Amalarethinam, "Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing" International Journal of Computer Application, 2011.

[10] Shahrzad Aslanzadeh, Venkatesh Mahadevan, Christopher Mcdermid, "Availability and Load Balancing in Cloud Computing" International Conference on Computer and Software Modeling IPCSIT, 2011.

[11] Poonam Devi & Mr. Trilok "GabImplementation of cloud computing by using short job scheduling" International Journal of Advanced Research in Computer Science and Software Engineering.

[12] Rana, Madhurima, S. Bilgaiyan, and U. Kar. "A study on load balancing in cloud computing environment using evolutionary and swarm-based algorithms." Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on. IEEE, 2014.

[13] Kousik Dasgupta, Brototi Mandal & Paramartha Dutta, "Load Balancing in Cloud Computing using Stochastic Hill Climbing-A Soft Computing Approach", Elsevier (C3IT) 2012.

[14] J. Uma, V. Ramasamy & A. Kaleeswaran, "Load Balancing Algorithms in Cloud Computing Environment-A Methodical Comparison", Vol.3, Issue.2,pp.272-274, 2014.

[15] Ms. Nitika, Ms. Shaveta, Mr. Gaurav Raj, "Comparative Analysis of Load Balancing Algorithms in Cloud Computing", International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May 2012.

[16] Suresh M, Ullah Z S, Kumar S, "An Analysis of Load Balancing in Cloud Computing", International Journal of Engineering Research & Technology 2: 428-433, 2013.l