Various Load Balancing Techniques in Cloud Computing: A Review

Anjuli Goel[1] and [2]Pooja Kalpesh

Chandigarh University
[1]*anjuli.e7948@cumail.in*
[2]*pooja.e7943@cumail.in*

Abstract

Cloud computing is getting popular day by day due to its ever increasing usage in various fields like metered bill, resource pooling, scalability, updated software's etc. and in turn there are new challenges are arising such as security of data, load balancing, quality of services, task scheduling etc. Load balancing is challenging issue due to ever increasing information on cloud. It is done by proper allocation of tasks and utilization of resources. Load balancing helps in reducing carbon emission and energy consumption. There are many load balancing algorithms available on cloud. All algorithms have their own way to work and have their pros and cons. This paper describes different types of load balancing methods with their comparison.

Keywords: *Cloud computing, Load balancing, services, algorithms*

1. Introduction

Cloud computing is a concept which includes a large number of devices connected with each other those can be accessed from anywhere at any time with use of internet [17]. The main objective of cloud computing is to provide services, do calculations and share data over a world-wide network [8]. It is the state-of-the-art in the world of computerised information. It has the advantages of previous known computing like distributed computing, grid computing and parallel computing. Cloud computing has many definitions according to its uses and users. One of the definitions is "a virtually interconnected collection of computers which can be added or removed systems dynamically and with less interference of service providers"[1]. It obliges changes sought after. Cloud is that technology which provides technology-enabled and on-demand services like server, applications, storage, network components and development and designing platforms and so on[21]. Cloud assets can be effectively scaled all over. Cloud assets are progressively designated to clients on interest. As the quantity of clients expands, the accessible assets decline progressively. Designation of cloud assets to client on interest offers ascend to the issue of load adjusting. On the off chance that outstanding task at hand isn't conveyed legitimately, a few hubs in cloud will be vigorously stacked and a few hubs will be under stacked. Similarly if the assets given by the cloud are not designated adequately, it prompts delay in giving support of the clients. Stack awkwardness may cause framework bottleneck. To accomplish asset usage and no postponement in giving administration asset designation ought to be done in a productive way.

This paper describes various load balancing algorithms. Rest of the paper is divided in to four sections: section 2 narrates about the taxonomy of cloud computing. Section 3 describes metrics of load balancing and load balancing algorithms. Section 4 draws the conclusion.

2. Taxonomy of cloud computing

2.1. Characteristics of cloud computing

In order to provide qualitative services, cloud has five essential features [5, 2, 17]

1. **On-demand self-service**: A client can avail computing devices like storage, server etc. automatically as required, even escalating service provider.
2. **Broad network access**: By using internet, one can access the cloud anywhere by the use of different devices such as mobile, laptops, tablets etc.
3. **Resource pooling**: The customer can access the pooled resources as a multi-tenant model provides by service provider which are dynamically provision and released as per requirements of customer.

4. **Rapid elasticity**: Customer can scale up the resources as needed and scale down the resources when finish the work easily.
5. **Measured service**: User has to pay for only those resources which user is availing. Cloud has a metering capability and user can pay according to usage of resources. It also known as pay as you go.

2.2 Service models of cloud

Cloud provides three types of services: Infrastructure as a service (Iaas), Platform as a service (Paas) and Software as a service (Saas) [3].

1. **Software as a service (Saas)**: The consumer can avail the service provider's infrastructure to host the applications with the use of internet. User can avail this service from anywhere in the world and requires no installation process. E.g. Salesform.com
2. **Platform as a service (Paas):** This service model provides the platform to user to host the applications. User can build its applications, deploy it and maintain it using platform layer. E.g. Azure, Heroku, Force.com.
3. **Infrastructure as a service (Iaas):** It provides framework like storage, servers, operating system etc. These resources can be accessed by users through virtual machines, instead of buying all the resources whenever required. E.g. Amazon EC2.

2.3 Deployment models of cloud computing

There are four types of deployment models on cloud, based on their access, size and ownership. These models are explained as follows [18]:

1. **Public cloud**: This model is open to use for general public. Any person can use the resources free or without any subscription. It is managed and maintained by organisation that provides this cloud model services.
2. **Private cloud**: When any organisation made its cloud and managed it, called private cloud. Only members of an organisation can use that cloud data, no one outside can access it. It provides better security and ease of maintenance.
3. **Community cloud**: In this model, many organisations that are belong from one community made their own cloud. An infrastructure is accessed, managed and maintained by the community members only.
4. **Hybrid cloud**: It is a blending of more than one clouds such as public and private. When an organisation use public cloud for non-critical tasks and use private cloud for some sensitive data then hybrid cloud exists.

3. Load Balancing

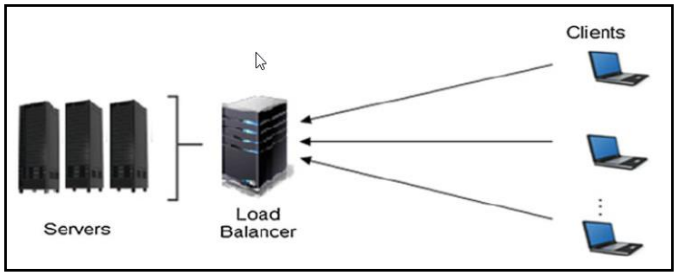

Figure 1. Diagram of Load Balancing [25]

3.1 Metrics for load balancing

Different metrics are taken into consideration for load balancing [8]:

1. **Scalability**: It is the ability of the system to use all the nodes uniformly in the network. It increases efficiency of the system.
2. **Resource Utilization**: Algorithms must have maximum utilization of the resources.

3. **Performance**: Performance of load balancing depends upon the above all parameters. If other parameters are working properly, then performance is good, otherwise not.
4. **Response time**: It is the time when an algorithm starts responding after submission. This respond time should be less.

3.2 Need of Load balancing

Load balancing is done for the avoidance of overloading and under loading of systems. It is not easy to be done on cloud computing environment because it also includes some more constraints with it like reliability, security, throughput etc. The main aim of load balancing algorithm is to minimize the response time by assigning the uniform load on several nodes. It is used to gain resource optimizations, i.e. to maximise utilization of resources & to minimise consumption of resources. It also makes sure that no one node is overloaded with the tasks. The load can be considered in terms of network load, delay memory usage and CPU load.

3.3 Categories of load balancing:

There are two types of load balancing algorithms in cloud environment [7].

1) **Static load balancing**: static algorithms evenly divide the load among the nodes. This type of load balancing requires prior information about the resources, so the assigning of load does not depend upon current system. It is useful in low level of load algorithms.
2) **Dynamic load balancing**: in this, under loaded system is searched and then assign the load on that device. It requires current state of the task to manage the resources. This is helpful for the real time communication network.

3.4 Static load balancing algorithms

a) Round-robin load balancing algorithm: The round-robin load equalisation algorithmic program uses the round-robin scheme for allotment of tasks [4][14]. It chooses the 1st device randomly, then allocate tasks to any or all other computing machines in an exceedingly spherical robin fashion. Processors unit of measurement assigned to each technique in an extremely circular order with none type of priority and so there isn't any starvation. This serves the advantage of quick response within the case of equal work distribution amongst processes [15]. However, completely different processes have different process times, therefore, at any instant of time few hubs could be stacked vigorously while remaining remain latent and under-used. [6]

b) Weighted Round-Robin Load Balancing Algorithm: Weighted round-robin was developed to improve the vital matters with spherical robin rule [4]. In weighted spherical robin algorithm, each server is assigned a weight and per the estimations of the loads, occupations square measure distributed. Processors with more noteworthy limits square measure doled out a bigger cost. Consequently the most noteworthy weighted servers can get extra errands amid a situation where all loads become equivalent, servers can get adjusted traffic.

c) Min-Min Load Balancing Algorithm: Min-min load equalization rule begins by finding the time for all undertakings. At that point among these base occasions, the base cost is choosing that will be that the base time among all errands on any asset [14]. As per that base time, the undertaking is then ordinary on the relating machine. The execution time for every extraordinary errand is refreshed consequently machine which task is expelled from the rundown. This system is pursued till the whole undertakings zone unit doled out the asset. In circumstances wherever the scope of little errands is extra than the scope of enormous assignments, this calculation accomplishes higher execution [4]. Be that as it may, this methodology will cause starvation [15].

d) Max-Min Load Balancing Algorithm: Max-min load compromise rule is like the min-min calculation with the exception of the accompanying: when searching for the base execution time, the most extreme cost is picked that will be that the greatest time among all errands on the assets [14].

At that point as per the most extreme time, the errand is booked on the comparing machine. The execution time for every single elective undertaking are refreshed subsequently machine and along these lines the assigned errand are expelled from the rundown of undertakings that are to be dispensed to the machines. Since the necessities are far-popular previously, this standard is relied upon to perform well.

e) Opportunistic Load Balancing Algorithm: Opportunistic load equalisation rule makes an attempt to stay every node busy [4]. Thus it will not contemplate the gift workload of every laptop. OLB dispatches unexecuted tasks too presently on the market nodes during a random order notwithstanding the node's current employment. Since OLB doesn't compute the execution time of the hub, the assignment to be prepared will be handled amid a slower way prompting bottlenecks in spite of a portion of the hubs being free [4].

3.5 Dynamic load balancing algorithms

a) Genetic Load Balancing Algorithm: This standard executes in powerful cloud setting and it utilized delicate registering approach. This standard is test from the normal improvement. This standard gives higher execution contrast with static algorithms rule. The upside of the rule is well handle a colossal inquiry house, relevant to cutting edge objective operate and will abstain from being tack into native optimum solution[11].The execution system of GA's is based on 3 stages.

- Choice Operator: Selects the underlying populace discretionarily.
- Hybrid Operator: See health consolidate of people for half and half.
- Change Operator: A tiny low or less possibility worth is named transformation worth.

These bits are flipped from 0s to 1s or 1s to 0s. The yield is new pool of people prepared for hybrid.

b) Game Theory Algorithm: This zestful algorithmic program operates publically cloud setting. This algorithmic program is partition the cloud into 3 categories specifically idle, traditional and overload supported the load degree. The overall population cloud incorporates a few hubs and it set at totally better places. This segment helps U.S.A. to deal with the huge cloud. The heap compromise began when the assigning, with the most controller choosing that cloud parcel should get the obligation and segment load balancer defines the obligation dispersion to the hubs. On the off chance that the cloud parcel is customary, at that point task total territorially and if the cloud segment load standing isn't conventional, this activity should be exchanged to an alternate. When the setting is enormous and exacerbates these divisions form increment the capability inside the open cloud setting. This heap compromise preparing the introduction and look after strength. The test of this algorithmic program is to anticipate the obligation entry, capacities of each hub inside the cloud [9, 10].

c) Stochastic Hill Climbing Algorithm: A variation of Hill mounting algorithmic program random Hill mounting and it provides the answer for optimisation downside. This method characterized into 2 strategies referred to as total and fragmented. Total system that ensures an explicit reply by two ways: by ensures that no such task exists and by finding a probably substantial task to the variable. On the contrary, deficient procedure doesn't ensure right responses for most of the given sources of info. A Hill mounting algorithmic program- random Hill mounting algorithm relies on the unfinished technique for resolution optimisation issues. SHC may be a native optimisation algorithmic program that ceaselessly moves within the upward direction for increasing the worth. In the event that no neighbour fuses a higher worth, at that point it'll precisely stop. This essential arrangement of this activity is rehash the appropriate response till found or halting no neighbour fuses a high worth. In this way it's 2 primary parts: a hopeful generator that maps one answer contender to a gathering of doable successors, and investigation standard that positions every substantial answer determined rising the examination winds up in higher arrangements[10, 12].

d) Ant Colony based Algorithm: The hymenopterous creepy crawly settlement algorithmic standard depends on the conduct of the genuine ants. The hymenopterous creepy crawly will see the ideal way wherever the source sustenance is advertised. The Ants while looking for a way from their settlement looking for sustenance discharge a compound called emission on the base along these lines going a way for different ants to pursue the trail. Anyway this substance vanishes with time. This methodology means to direct efficient conveyance of work among the hub. The hymenopterous creepy crawly keeps up the record of each visited hub for higher making bring in future. The hymenopterous creepy crawly would store pheromones all through their development for option hymenopterous bug to pick next hub. The force of pheromones will fluctuate on the bases of beyond any doubt factors like separation of nourishment, nature of sustenance and so forth once the errand gets blessed the pheromones is refreshed. Each hymenopterous creepy crawly creature collect their own one of a kind individual result set and it's some time later incorporated with an absolute answer. The hymenopterous bug unendingly invigorates one outcome set as opposed to change their own one of a kind result set. By the underground creepy crawly pheromones fundamentals, the proper reaction set is always revived [13, 16].

e) Firefly Algorithm: The dynamic firefly arranging algorithmic guideline is relies upon glimmering qualities of fireflies and it is applicable in streamlining the calendar technique to the cloud organize. The methodology included concerning business balance, accordingly they utilized after 3 leads inside the cloud framework [19].

- All fireflies are hermaphroditic all together that one firefly will pulled in to various fireflies.
- Engaging quality is relative to their spend or, along these lines for any 2 blazing fireflies, the less splendid firefly can move towards the more brilliant one. The appeal is corresponding to the splendor and that each abatement for the separation increments.
- The brilliance of this algorithm is influenced and controlled by the aim perform.

f) HBB-LB Algorithm: This algorithm is inspired from nature algorithmic standard for association. Honey Bees comprises a ruler and seekers wherever seekers are categorizes into two; utilized and out of work. The seekers are privy with respect to the sustenance available near to waggle move by scout honey bee (jobless), the move is explain about the data to the contrary search honey bees in regards to the separation, quality, course and elective information that is valuable in getting the food[20]. This algorithmic standard has comparative guideline in parity crafted by the vms. This algorithmic standard ascertains the computing devices work, afterwards it chooses which node is over loaded, lightweight weighted or adjusted. The HBB algorithmic principle figures the virtual machine work then it chooses whether or not it's over loaded, lightweight weighted or adjusted. These undertakings are alluded to as scout honey bee inside the subsequent stage. Bumble bee behaviour inspired Load levelling method decreases the inertness of VM and also diminishes the holding up time of undertaking [22, 23, 24].

4. CONCLUSION

Load balancing algorithms are very crucial for efficient and maximum utilisation of resources. Some systems are overloaded and some are under-utilized. So, different types of algorithms have been implemented with different parameters. This paper describes the working of various currently applicable static and dynamic load balancing algorithms. Every algorithm has its pros and cons. In future works, algorithms can be made with fault tolerance parameter utilisation of resources, less respond time and maximum throughput in the cloud computing environment.

References

1. "NIST Definition of cloud computing," http://www.nist.gov./itl/cloud/upload/cloud-def-v15.pdf
2. http://www.trainingindustry.com/media/3956976/gh%20cloud%20computing%20characteristics%20are%20key.pdf
3. "Introduction to Cloud Computing," http://www.dialogic.com/~/media/ products/docs/whitepapers/12023-cloud-computing-wp.pdf, July 2010.
4. "Cloud Load Balancing Image,". https://www.xcellhost.cloud/blog/importance-load-balancing-cloud-computing-environment [Online]

5. Tharam Dillon, Chen Wu and Elizabeth Chang," Cloud Computing: Issues and Challenges," Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications, DOI 10.1109/AINA.2010.187, 1550-445X/10 2010 IEEE.
6. Cloud computing services, available at: https://www.rackspace.co.uk
7. Sajjan R.S and Biradar Rekha Yashwantrao," Load Balancing and its Algorithms in Cloud Computing: A survey" JCSE International Journal of Computer Science and Engineering, pp. 95 – 100, 2017.
8. Deepak B S, Shashikala S V and Radhika K R," Load Balancing Techniques in Cloud Computing: A Study" International Conference on Information and Communication Technologies, pp. 1 – 4, 2014.
9. Xu G, PangJ and Fu X," A load balancing model based on cloud partitioning for the public cloud" Tsinghua Science and Technology, 2013.
10. Mondal B, Dasgupta K, and Dutta P.," Load balancing in cloud computing using stochastic hill climbing – A soft computing approach," Procedia technology, 2012.
11. Dasgupta K, Mandal B, Dutta P, Mandal JK and Dam S. A," Genetic Algorithm based load balancing strategy for cloud computing" Procedia technology, pp. 1- 7, 2013.
12. Liu G, Li J and Xu J," An improved min-min algorithm in cloud computing," Proceedings of the 2012 International Conference of modern Computer Science and Applications, Springer, pp. 47 – 52, 2013.
13. Mishra R and Jaiswal A," Ant colony optimization: A solution of load balancing in cloud," International journal of Web and Semantic Technology, 2012.
14. T. Desai and J. Prajapati, "A Survey of Various Load Balancing Techniques and Challenges in Cloud Computing," vol. 2, issue 11, pp. 158-161.
15. Shiny, "Load Balancing in Cloud Computing: A Review," vol. 15, issue. 2, pp. 22-29, 2013.
16. Sakthipriya N and Kalaipriyan T," Variants of ant colony optimization –a state of an art," Indian journal of Science and Technology, 2015.
17. Cloud computing, available at :http://en.wikipedia.org/wiki/Cloud_computing
18. http://www.en.wikipedia.org/wiki/Cloud_computing#Deployment_models
19. Kokilavani T, and Amalarethinam DD," Load balancing min-min algorithm for static meta-task scheduling in grid computing." International journal of Computer Applications, 2011.
20. LD DB and Krishna PV," Honey bee behaviour inspired load balancing of tasks in cloud computing environments," Applied soft computing,2013.
21. M. Sajid and Z. Raza, "Cloud computing: Issues and Challenges," International Conference on Cloud , Big Data and Trust, pp. 35 – 41, 2013.
22. Anju Baby J," A survey on honey bee inspired load balancing of tasks in cloud computing," International journal of Engineering Research and Technology, 2013.
23. Nakrani S and Tovey C," On honey bees and dynamic server allocation in internet hosting centres," Adaptive Behaviour, 2004
24. Chakaravarthy T and Kalyani K," A brief survey of honey bee mating optimization algorithm to efficient data clustering," Indian journal of Science and Technology, 2015.
25. R.Z. Khan and M.O. Ahmad," Load balancing Challenges in Cloud Computing: A survey," Proceedings of the International Conference on Signal, Networks, Computing, and Systems, pp. 25 – 32, 2016.