Review Paper: A Novel Cloud Framework for Efficient and Effective Load Balancing

Tarandeep Kaur[1], Navneet Kaur Sandhu[2]
[1,2]*Dept. of Computer Engineering*
[1,2]*Desh Bhagat University*,Mandi Gobindgarh, India
[1]tarandeepbhandol93@gmail.com

Abstract— Cloud Computing is one of the very useful and emerging area in the field of IT environment. An important role is provided in cloud computing environment is Load Balancing. The scheme of Efficient load balancing ensures efficient resource utilization by way of provisioning of resources to cloud user's on-demand basis in pay-as-you-use-manner. A very useful scheduling criteria will be applying to help the customers to select the best load balancing criteria. In this paper we provide various load balancing schemes in distinct cloud surroundings based totally on requirements specific in-Service Level Agreement (SLA).

Keywords—Cloud Computing, Load Balancing, Resource Provisioning, Resource Scheduling, Service Level Agreement (SLA)

I. Introduction

A. *Cloud computing:-*

The cloud computing is one of the most trending applied sciences in IT domain. It is a technique of coping with and pooling services like servers, records base, storage, software and greater over the internet based totally on the user's want or demand. Users can get the sources from the data facilities as per their requirements from somewhere thru a net linked computer or hand held devices. One of the challenging tasks in cloud computing is load balancing used to allocate work load among the statistics centers. Datacenters are physical machines that has the responsibility to complete the request and demand of cloud users. So, load balancing is required to manipulate the load throughout data centers, limit the overload, enhance performance, minimize common execution time and supply higher aid utilization. Load balancing can decrease the response time and maximize the user's satisfaction. It also increases the source utilization and restriction the energy consumption.

Types of cloud providers

- **Software as a Service (SaaS): -** SaaS clients hire utilization of functions jogging inside the Cloud's issuer infrastructure, for example SalesForce. The functions are generally presented to the customers by using the Internet and are managed totally through the Cloud provider. That capacity that the administration of these offerings such as updating and patching are in the provider's responsibility. One huge gain of SaaS is that all clients are running the equal software program version and new functionality can be effortlessly integrated through the company and is consequently accessible to all clients.
- **Platform as a Service (PaaS)**: - PaaS Cloud providers offer a software platform as a service, for instance Google App Engine. This permits customers to install custom software the use of the equipment and programming languages presented by the provider. Clients have manipulated over the deployed functions and environment-related settings. As with SaaS, the management of the underlying infrastructure lies inside the accountability of the provider.
- **Infrastructure as a Service (IaaS)**: **-** IaaS grants hardware assets such as CPU, disk space or community elements as a service. These sources are commonly delivered as a virtualization platform by the Cloud provider and can be accessed throughout the Internet by the client. The client has full control of the virtualized platform and is no longer accountable for managing the underlying infrastructure.

How Cloud Computing Works

Let's say you are an executive at a large corporation. Your specific duties encompass making positive that all of your personnel have the right hardware and software program they need to do their jobs. Buying computer systems for anyone is not sufficient -- you additionally have to buy software or software program licenses to give personnel the equipment they require. Whenever you have a new hire, you have to buy greater software or make certain your cutting-edge software program license permits every other user. It's so annoying that you locate it challenging to go to sleep on your large pile of cash every night. Soon, there might also be an choice for executives like you. Instead of putting in a suite of software program for every computer, you'll solely have to load one application. That software would allow employees to log into a Web-based carrier which hosts all the programs the consumer would need for his or her job. Remote machines owned with the aid of some other organization would run the entirety from e mail to phrase processing to complex facts analysis programs. It's known as cloud computing, and it ought to change the whole laptop industry. In a cloud computing system, there is a vast workload shift [5]. Local computer systems no longer have to do all the heavy lifting when it comes to walking applications. The community of computer systems that make up the cloud handles them instead.

B. Introduction to load balancing:-

Load Balancing is the method of improving the overall performance of the machine by means of transferring of workload among the processors. Workload of a laptop means the total processing time it requires to execute all the tasks assigned to the machine. Balancing the load of digital machines uniformly capacity that everybody of the accessible laptop is no longer idle or partly loaded whilst others are closely loaded. Load balancing is one of the necessary factors to heighten the working performance of the cloud service provider. The benefits of distributing the workload consists of multiplied useful resource utilization ratio which in addition leads to bettering the common performance thereby accomplishing maximum consumer satisfaction. In cloud computing, if users are increasing load will also be increased, the extend in the variety of customers will lead to poor performance in terms of useful resource usage, if the cloud company is no longer configured with any desirable mechanism for load balancing and also the potential of cloud servers would no longer be utilized properly.

The primary goals of load balancing are mentioned below:

- To treat all jobs in the system equally regardless of their origin.
- To improve the performance of the cloud substantially.
- To maintain the system stability.
- To maximum throughput, minimize response time, and avoiding overload.
- To have a backup plan in case the system fails even partially.
- To provide optimal resource utilization.
- To accommodate future modification/changes in the system.

II. LOAD BALANCING ALGORITHMS REVIEW

In this section we talk about the most known contributions in the literature for load balancing in Cloud Computing. We classify the load balancing algorithms into two types:

- Static algorithms
- Dynamic algorithms.

We first discuss the static load-balancing algorithms that have been developed for Cloud Computing. Then, we will discuss the dynamic load-balancing algorithms.

A. Static Load Balancing Algorithms

In this scenario the prior knowledge of resources such as capacity, processing power, number etc. is required. Any change in load required at run time is not possible. Though it is easy to implement but not much suited for various cloud environments especially where it is not possible to fix the requirements and resources. Static load balancing algorithm is basically assigning works to nodes primarily based solely

Static Load balancing algorithms dole out the obligations to the nodes essentially dependent on the capability of the node to manage new tasks. The framework is put together absolutely completely with respect to earlier comprehension of the hubs' properties and capacities. These would incorporate the hub's handling memory and capacity limit, and most current perceived verbal trade execution. Despite the fact that they may moreover incorporate skill of the verbal trade earlier execution, static calculations ordinarily don't consider dynamic changes of these qualities at run-time. In addition, these algorithms can't adjust the load changes for run-time.

Etminani proposed another plan of load balancing dependent on two methods Min-Min and Max-Min by utilizing their points of interest and attempted to diminish the finish time. They have known as it min-max min-min selective. They assessed their experiment with Gridsim in static condition.

M. Randles done comparative study for cloud computing on distributed load balancing algorithm. They said that there are three methods for large scale load balancing in cloud systems-

- Random sampling of system domain,
- Restructured system to optimize job assignments
- Nature inspired

M. Nakai proposed Server-based load adjusting for Internet appropriated administrations, which is a load balancing approach for web servers apportioned on huge scale. In this plan the specialist attempted to limit the reaction time by methods for applying the cutoff points on redirection of requests to various remote servers.

Junjie proposed a load adjusting algorithm [13] for the private Cloud the utilization of virtual registering device to real machine mapping. A central scheduling controller and a resource monitor of the algorithm are included in the structure. The planning/scheduling controller does practically everything for computing which valuable asset can take the task and after that relegating the test to that particular resource. However, the

resource monitor carries out the responsibility of gathering the insights regarding the asset's accessibility. The way toward mapping undertakings goes by means of four basic stages which are:
accepting the virtual machine demand, at that point getting the sources subtleties utilizing the resource monitor. From that point onward, the controller ascertains the assets capacity to deal with errands and the resources that gets the best evaluating is the one accepting the undertaking. At last, the customer will most likely access the application.

The proposed algorithm in the [11] is an expansion to the MapReduce algorithm [12]. MapReduce is a model which has two key assignments:

It Maps obligations and Reduces undertakings results. Additionally, there are three techniques in this model. The three techniques are group, part and comp. MapReduce first execute the segment way to deal with incite the Mapping of tasks. At this stage the request is parceled into segments utilizing the Map tasks.

At that point, the key of each part is spared into a hash key work area and the comp system does the examination between the parts. From that point forward, the team system associations the pieces of practically identical substances the utilization of the Reduce tasks. Since different Map tasks can think about substances in parallel and procedure them, this will be thought process the Reduce undertakings to be over-loaded. In this way, it is proposed in this paper to include one additional heap adjusting stage between the Map task and the Reduce task to minimize the over-loaded on these assignments.

The load balancing in the middle partitions exclusively the enormous tasks into small tasks and afterward the small tasks are dispatched to the Reduce tasks fundamentally dependent on their accessibility.

B. Dynamic Load Balancing Algorithms

Dynamic load balancing algorithms consider the various qualities of the nodes capacities and system data transfer capacity. The majority of these algorithms rely upon a mix of knowledge basically dependent on earlier assembled realities about the nodes in the Cloud and run-time properties accumulated as the chose node framework the assignment's segments. These algorithms relegate the undertakings and can likewise powerfully reassign them to the nodes principally dependent on the properties assembled and determined. Such algorithms require ordinary checking of the nodes and challenge progress and are regularly increasingly hard to execute. In any case, they are progressively right and might need to final product in increasingly proficient load balancing.

This is performed dependent on a complexity of the SUM of associations of every node in the Cloud and afterward the task is assigned to the node with least variety of connections.
However, WLC does now not take over the abilities of each node, for example, preparing storage limit, data transfer capacity and speed. The proposed algorithm is called ESWLC (Exponential Smooth Forecast dependent on Weighted Least Connection). ESWLC improves WLC by method for considering the time succession and trails. That is ESWLC assembles the decision of allocating a beyond any doubt challenge to a hub in the wake of having various obligations appointed to that hub and getting the opportunity to perceive the node abilities. ESWLC assembles the decision essentially dependent on the experience of the node's CPU control, memory, amount of associations and the amount of space by and by being utilized. ESWLC then predicts which node is to be chosen dependent on exponential smoothing.

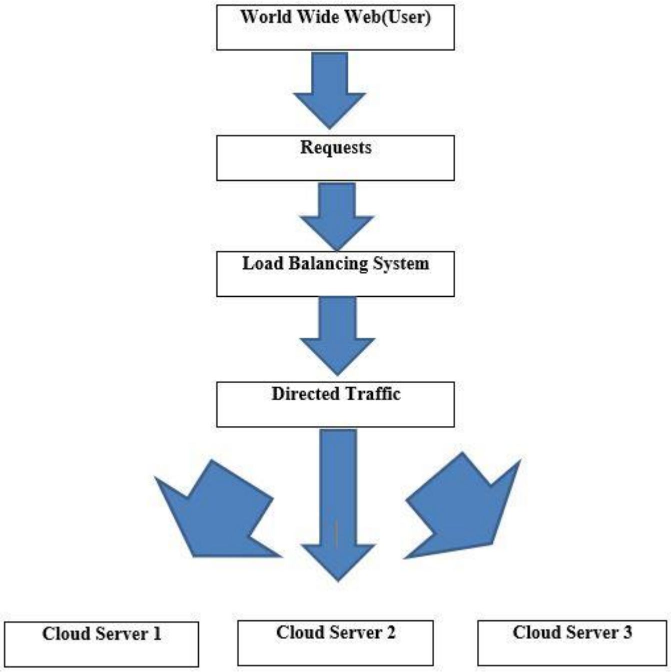
Nitish C. proposed load balancing scheme HEFT based workflow scheduling for cost Optimization with in Deadline in Hybrid clouds. They have simulated their work on workflowsim. They have taken deadline completion as essential difficulty in load balancing for impartial duties and tried to decrease the ordinary cost.

The algorithm proposed in [17] is a twin direction downloading algorithm from FTP servers (DDFTP). The algorithm presented can be also implemented for Cloud Computing load balancing. DDFTP works by means of splitting a file of size m into m/2 partitions. Then, every server node starts processing the undertaking assigned for it based on a sure pattern.

Y Lua proposed a load balancing plan called as Join-Idle-Queue. This plan is fundamentally founded on administered load adjusting which is finished by utilizing dispensed dispatcher. At the first step every circulated dispatcher makes inert processors line. At that point join these inactive lines to dole out the occupations going to the servers to limit the heap of other over-burden nodes. It also claims to confine the response time.

We furthermore look at these algorithms basically dependent on the difficulties referenced in Section II. As referenced before, the outstanding strategies give one of a kind choice to load balancing that swimsuit a few circumstances anyway not others. The static algorithms are generally effective in expressions of overhead as they do now not have to monitor the benefits for the length of run-time. In this way, they would work very well in a relentless situation where operational properties do never again change after some time and loads are commonly uniform and consistent. The dynamic algorithms then again given by the drove better arrangement that should manage the load powerfully at run-time basically dependent on the discovered homes of the assets at run time.

However, this function prompts high overhead on the machine as relentless observing and oversee will include more noteworthy guests and may also reason additional delays. Some recently proposed dynamic load balancing algorithms attempts to avoid this overhead by method for using novel undertaking appropriation models.



Purposed Model

- User (WWW): - It is the first input of our proposed algorithm. In this the user can upload or send all the requests to server via any browser.

- Request: - Request is the 2nd parameter considered in proposed algorithm where all the requests are sending to server.

- Load Balancing System: - Load balancing system is one of the very powerful and intelligent system where the load balancing system/algorithm can manage/arrange all the requests that are coming from the browser. At, this level all the coming requests are properly accepted and make a temporary table of all upcoming requests or work load.

- Directed Traffic: - At this level all the requests are divided according to the work load, efficiency, free resources of the server.

Algorithm	Dynamic Environment	Static Environment	Distributed Balancing	Hierarchical Balancing	Centralized Balancing
Max Min	No	Yes	No	No	Yes
Ant Colony	Yes	No	Yes	No	No
CLBDM	No	Yes	No	No	Yes
Particle Swarm Optimization	Yes	No	Yes	No	No
Biased Random Sampling	Yes	No	Yes	No	No
Map Reduce	No	Yes	Yes	Yes	No
Min Min	No	Yes	No	No	Yes
LBMM	Yes	No	No	Yes	No
Active Clustering	Yes	No	Yes	No	No

III. CONCLUSION AND FUTURE SCOPE

Load Balancing is a fundamental task in Cloud Computing condition to increase the usage of resources. In this paper, we referenced various load balancing plans, each having a few upsides and downsides. On one hand static load adjusting plan give simplest observation and simulation of condition anyway neglect to show heterogeneous nature of cloud. Then again, dynamic load balancing algorithms are hard to simulate however are fantastic satisfactory in heterogeneous condition of cloud computing. Additionally, the degree at node which actualizes this static and dynamic algorithm plays out a vital job in choosing the viability of algorithm. In contrast to incorporated calculation, allocated nature of hierarchy gives better adaptation to non-critical failure anyway requires higher level of replication and on the distinctive hand, progressive algorithm isolate the heap at special phases of pecking order with upper dimension nodes mentioning for administrations of lower degree nodes in adjusted way. Thus, dynamic load adjusting strategies in dispensed or various leveled condition give higher execution. In any case, execution of the distributed computing environment can be also expanded if conditions between duties are modeled displayed using workflows.

IV. REFERENCES

1. Chitra DD, Uthariaraj VR. Load balancing in cloud computing environment using Improved Weighted Round Robin Algorithm for non-preemptive dependent tasks. The Scientific World Journal. 2016.
2. Solmaz A, Motamedi S, Sharifian S. Task scheduling using Modified PSO Algorithm in cloud computing environment. International Conference on Machine Learning, Electrical and Mechanical Engineering; 2014. p. 37–41.
3. Imran MA, Pandey M, Rautaray SS. A proposal of resource allocation management for cloud computing. International Journal of Cloud Computing and Services Science. 2014; 3(2):79–86.
4. Jamuna RMR, Gouda KC, Nirmala N. Load balancing technique for climate data analysis in cloud computing environment. International Journal of Science, Engineering and Computer Technology. 2013; 3(5):183–85.
5. Namrata G, Garala K, Maheta P. Cloud load balancing based on ant colony optimization algorithm. IOSR Journal of Computer Engineering (IOSR-JCE); 2015. p. 11–18.
6. Danilo A. Quality-of-service in cloud computing: modeling techniques and their applications. Journal of Internet Services and Applications. 2014; 5(1):1–17.
7. Jia Z. A Heuristic clustering-based task deployment approach for load balancing using Bayes Theorem in cloud environment. IEEE Transactions on Parallel and Distributed Systems. 2016; 27(2):305–16.
8. Kunjal G, Goswami N, Maheta ND. A performance analysis of load Balancing algorithms in Cloud environment. 2015 International Conference on Computer Communication and Informatics (ICCCI), IEEE; 2015. p. 4–9.
9. Beghdad BK, Benhammadi F, Benaissa F. Balancing heuristic for independent task scheduling in cloud computing. 2015 12th International Symposium on Programming and Systems (ISPS), IEEE; 2015.
10. Aditi S, Sharma S. Credit based scheduling using deadline in cloud computing environment. International Conference on Resent Innovation in Science Engineering and Management; 2016. p. 208–16.
11. Sukhjinder GS, Vivek T. Implementation of a hybrid load balancing algorithm for cloud computing. International Conference on Science, Technology and Management; 2016. p. 173–82.
12. Mohana PS, Subramani B. A new approach for load balancing in cloud computing. International Journal of Engineering and Computer Science. 2013.
13. Shreya S, Kaur A. Load balancing in cloud computing using Shortest Job First and Round Robin Approach. International Journal of Science and Research. 2015; 9(4):1577–80.
14. Divya C, Chhillar RS. A new load balancing technique for virtual machine cloud computing environment. International Journal of Computer Applications. 2013; 69(23):37–40.
15. Yang X, HongTao L. Load balancing of virtual machines in cloud computing environment using improved ant colony algorithm. International Journal of Grid and Distributed Computing. 2015; 8(6):19–30.
16. Abbas RH, Katti CP, Saxena CP. A load balancing strategy for Cloud Computing environment. 2014 International Conference on Signal Propagation and Computer Technology (ICSPCT), IEEE; 2014.
17. Babu DL, Venkata PK. Honey bee behavior inspired load balancing of tasks in cloud computing environments. Applied Soft Computing. 2013; 13(5):2292–303.
18. Elhossiny I, El-Bahnasawy N, Omara FA. Job scheduling based on harmonization between the requested and available processing power in the cloud computing environment. International Journal of Computer Applications. 2015; 125(13):1–4.
19. Elrasheed I, Alamri F. Optimized load balancing based task scheduling in cloud environment. International Journal of Computer Applications; 2014.p. 35–8.
20. Ali A, Omara FA. Task scheduling using hybrid algorithm in cloud computing environments. IOSR Journal of Computer Engineering. 2015; 17(3):96–106.
21. Sourav B. Development and analysis of a new cloudlet allocation strategy for QoS improvement in cloud. Arabian Journal for Science and Engineering. 2015; 40(5):1409–25.
22. Nizomiddin BK, Choe TY. Dynamic task scheduling algorithm based on ant colony scheme. International Journal of Engineering and Technology (IJET). 2015; 7(4):1163–72.