Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

51

Load Balancing Techniques in Cloud Computing

Neha Bhosale

nehabhosale146@gmail.com

School of Computer Science, MIT WPU Pune

Tejashri Shinde

tejashrishinde31@gmail.com

School of Computer Science, MIT WPU Pune

Shantanu Nimbalkar

shantanunimbalkar09@gmail.com

School of Computer Science, MIT WPU Pune

Prof. Devyani Kambale

School Of Computer Science, MIT-WPU Pune

Abstract

A robust concept of cloud computing enables individuals and businesses to get the necessary services in accordance with their environment requirement. The model offers an array of functionalities which contain storage, deployment platform, easy to access resources on web. In the cloud, load balancing a frequent complication that makes it difficult to maintain performance with respect to service level Agreement (SLA) and be close to quality of Service (QOS)measurement contract as needed by the cloud service providers to organizations. To distribute a similar workload distribution across servers is a obstacle for

cloud service providers. To solve this issue, a concept known as load balancing strategies that improve network management conformance was recently put out. Due to needs fulfilment and resource restrictions on the network, load balancing becomes essential in order to distribute traffic across available resources more effectively and reliably.

Keywords: Load Balancing, Cloud Computing, Algorithms, services, Load Balancing Algorithm (LBA).

1. Introduction

A well-known technology called cloud computing offers services (both private and public), including scalable online storage services in place of locally stored information on users' devices such PCs or phones, and convenient internet-based access to data, programmes, and files (cloud).

In Fig. A below, a compilation of cloud computing is working is presented. To direct the cloud environment which is very crucial job, all cloud entities cooperate. For instance, by confirming that the assistance offered by CSPs are of the greatest calibre and honesty, cloud auditors act as the police of the cloud. In cloud environment there is Cloud carriers which make sure that there is a strong connection between user and cloud. The data center is located within the company's network in a private cloud, while in a public cloud it is located online or maybe It can be handled by cloud service providers (CSPs), and in a hybrid cloud it may be located in both.

Now we gonna discuss about infrastructure.so basically cloud infrastructure can be divided into two parts: the frontend side which user sees and the backend side which csp's handle. For better idea about this is mentioned below in Fig. B. The application dynamically schedules incoming user requests, and then allocates resources to clients using virtualization. The cloud's dynamic resources are managed using the virtualization technology, which is also utilised to balance the system's load.

Users submit requests through the internet, which are then stored in virtual machines (VMS).

CSPs mostly care about the user's time, in simple words they believe in quality of service which means users one or many requests should be treated on priority and finished in minimum time frame. So how its done, users request are given to the virtual machine using

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

scheduling techniques, which will eventually distribute the many user requests on different servers. But if there are many requests on cloud server it is hard to distribute these requests because every VM Is already working on something to overcome this problem to utilize the resources fully CSP should consider dynamic load balancer.

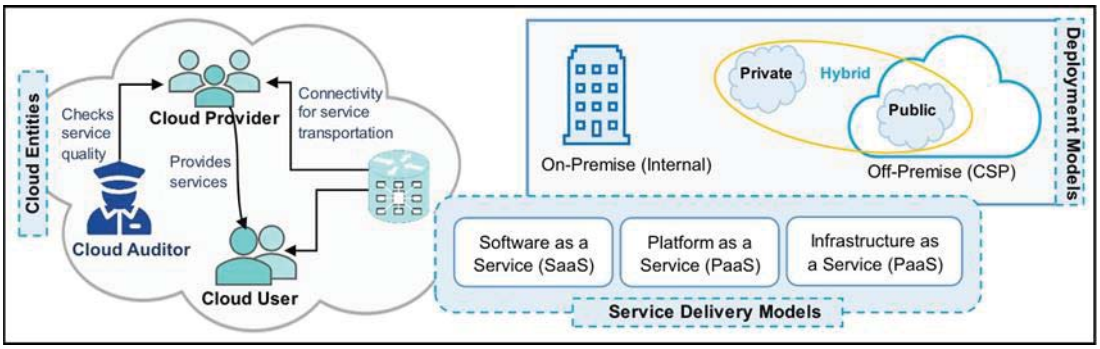


Fig A

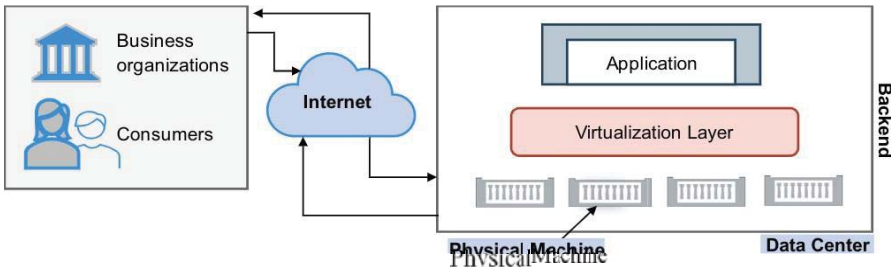


Fig B

The main problems of current systems are their high cost, lengthy maximum execution times, and excessive power consumption from overloaded Processors. The main objective is to create a framework for dynamic virtual server deployment of a resource's webserver in the cloud. Achieving this goal requires proper resource supply, as well as dynamic virtual machine deployment that lowers overall energy use.

The dynamic allocation of workload is a significant issue in cloud computing. The entire processing time needed to execute all of the tasks entrusted to the machine constitutes its workload. System performance is improved by load balancing by dividing the task among the processors. Workload benefits include enhanced overall performance and maximum client satisfaction, both of which are influenced by resource efficiency. Throughput is boosted and

reaction time is slashed in a cloud system by evenly dividing the total amount of virtual machines' workload.

2. Literature Review

We are going to discuss the load balancing related concepts which were already researched by many people, we are going to review those papers and share some thoughts about it. We'll go over the concept of load balancing, paying particular attention to its model, metrics, and currently popular solutions. leading to current load balancing research, where the suggested approaches of academics are discussed and evaluated. New algorithms that current researchers in the field of load balancing have proposed are then presented.

2.1 Load Balancing

Why we need Load Balancing, in shortly if there are many requests on cloud server (for eg. Many users are watching same cricket match on jiotv), which doesn't contain any scheduling strategy or dynamic load balancer. So, it makes the functioning of server very slow, that's why we need load balancing.

By implementing load balancing, the VMS resources can be optimised in the Cloud Computing environment. In cloud computing environment because of its raising usage Load balancing now becoming crucial technique for assuring a fair and dynamic task distribution and efficient resource utilisation. Improved resource allocation and increased user satisfaction are the benefits of a more efficient task distribution. Most crucial outcome of using load balancing in cloud systems is that it reduces lag times in data transmission and reception and prevents the situation where one machine or particular serve contains more requests than its processing capacity which will affect the QoS of cloud data centers.

2.2.1 Load balancing Model

How can we implement Load balancing in cloud environment, so there is a model already defined by researchers which CSP's uses to enhance their cloud environment.

In Fig. C below you can get the idea of how does load balancing works. The user request is analysed and transmitted to the selected Data Centre based on the resources that are available. Both of the servers must receive a fair proportion of the workload; neither should be

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

overloaded nor underloaded. This is when load balancing comes in handy. It is necessary to have a reliable load balancer in place to maintain the performance of cloud applications. Task Scheduling is one of many potential causes of uneven load distribution. The resources won't be used effectively if tasks aren't scheduled properly. The backside of the cloud is where load balancing takes place.

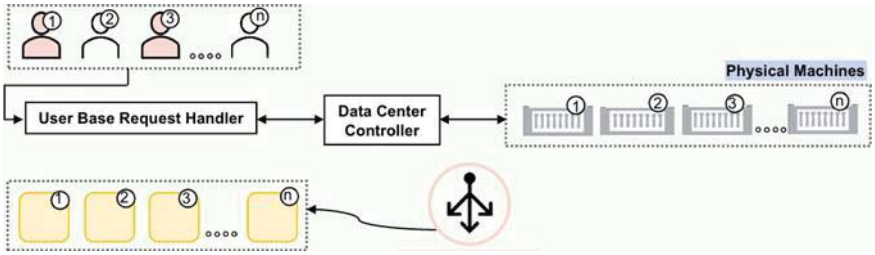


Fig C

2.2.2 Load Balancing metrics

After Model there are some protocols which make sure that the given load balancing technique is beneficial for cloud environment or not. We call these protocols a metrics. When creating and constructing a load balancing algorithm, these metrics are crucial.

- Resource Utilization (RU): Server utilization measures the percentage of resources that are being used on each server. Load balancing algorithms should always look to distribute the workload on every node of the server which will eventually decrease the processing time and optimize the resource utilization.
- Scalability(S): As a name scalability it calculates that how much scalable is particular algorithm. Load balancing algorithms should be scalable to accommodate the growth of the system and support the addition of new servers.
- Throughput (TP): Throughput works like quantity manager how much quantity of requests can particular server handle is checked by Throughput. Load balancing algorithms should aim to maximize throughput by distributing traffic evenly among servers and ensuring that no server is underutilized.
- Response Time (RT): Response time checks the time taken by particular serve to respond to specific requests. Load balancing algorithms should strive to minimize

response time by distributing traffic efficiently and ensuring that servers are not overloaded.

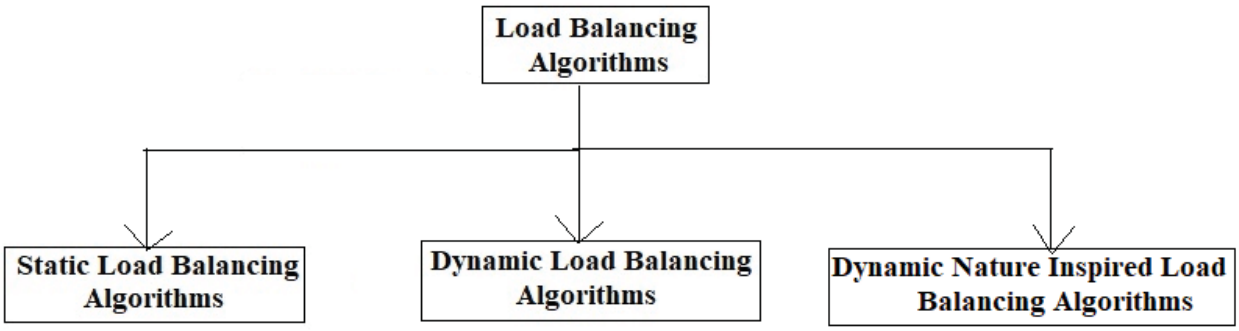
- Makespan (MS): Makespan checks the whole time taken by server to complete the one request. It is very important metrics because if server taking to much time, we can work on resource utilization of server. Lesser makespan the good is the load balancing algorithm (it works in inversely proportional manner).
- Fault Tolerence (FT): In case if server fails in middle of execution there should be a backup which will save the data that's what fault tolerance do it will continue the functioning after server fails. It is achieved through the use of redundancy and failover mechanisms that redirect traffic to available servers. Fault tolerance ensures high availability and performance of cloud-based applications and is an important consideration in load balancing design.
- Assoiciated Overhead (AO): The additional workload that the load balancing method produced. Task migration and inter-process communication may have played a significant role. When the system's load is balanced, the load balancing algorithm has the lowest overhead.
- SM Violation: it shows the amount of SLA violations has been done in load balancing.
- Migration Time (MT): as the name suggests it calculates the time taken to migrate between virtual machines

2.2.3. Current Standard Load Balancing Techniques

There are Many load balancing techniques, which is researched and developed by great minds so in this point we sill describe how load balancing strategies are categorized. Due to a number of restrictions, which contain overcrowding of nodes (many requests were send to the particular node), which makes a situation where only single node is working and other nodes are empty, there are many static algorithms which was initially very good to use when there was no that much crowd on cloud but they are still used in cloud environment very frequently, just like Round-Robin but it is not efficient today because of load of data Then again, the usage of context switching is the outcome of static quantum, which delays

processing and results in the rejection of jobs. This results in uneven workload distribution and inappropriate task assignment.

In Fig. D below. This diagram's function is to categorizes algorithms according to the fundamental methodology employed in this review study. To improve the situation in cloud computing environment to improve its performance we use cloud computing algorithms. They are mainly divided in three categories underlying their environments: nature-inspired algorithms, static algorithm and, dynamic algorithm., which are covered here.



1. Static Load Balancing (SLB) Algorithms: As a name suggests static which means it doesn't change for anyone in simple words it will not change its scheduling strategy or any other things with respect to Requests it will stick to its core functionality. It How it works. it is a technique that distributes incoming traffic across a fixed set of resources or servers. The distribution is determined at system setup or configuration and remains static until the system is reconfigured. Static load balancing can be simpler to implement and maintain compared to any load balancing techniques, but may not be as effective in handling fluctuating traffic patterns
2. Dynamic load balancing (DLB): As a name suggests Dynamic it is very different from static it doesn't stick to its core functionality it changes as per the request pattern to enhance the performance and resource utilization. How it works, it is a technique that adjusts the distribution of traffic across multiple servers or resources based on real-time traffic and resource utilization metrics. The distribution can change dynamically to ensure that resources are optimally utilized and traffic is evenly distributed, which will eventually enhance the performance of the server. Dynamic

load balancing can handle fluctuating traffic patterns more effectively than static load balancing.

3. NLB algorithms (Nature-inspired load balancing): As a name suggests nature inspired so it will make changes which is beneficial for system to enhance the performance, it follows a phenomenon to enhance the performance. How it works, it is a technique that works based on phenomena. These techniques can be used to optimize the distribution of traffic across multiple servers or resources and which will improve the overall performance. This type of technique is used to handle complex problems which was not handled by traditional techniques.

2.3 The most recent research on load balancing

We have taken look in some of the load balancing algorithms which were used by cloud service providers more frequently. but lack something. The following strategies offer effective load balancing techniques in an effort to enhance cloud computing performance. The algorithms' merits and shortcomings are discussed.

2.3.1. Min-Min algorithm [20]

Min-Min (MM) [1]: This method takes into account the shortest amount of time which was taken by any task with respect to its scheduling. MM has a number of drawbacks, such as the inability to run multiple tasks at once and the algorithm's high priority for smaller activities which starves larger processes and causes an imbalanced VM load. The suggested technique offers a matrix for task storage. It assigns tasks to its resource (VM) while accounting for execution and completion times. Jobs are scheduled based on two criteria: first criteria are to take into account the shortest time taken by any task on the VM and expected minimum execution time. After the task completion it does not look for current load of the VMS and updated load of the VMS in the process of task allocation, which is a drawback of the suggested solution.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

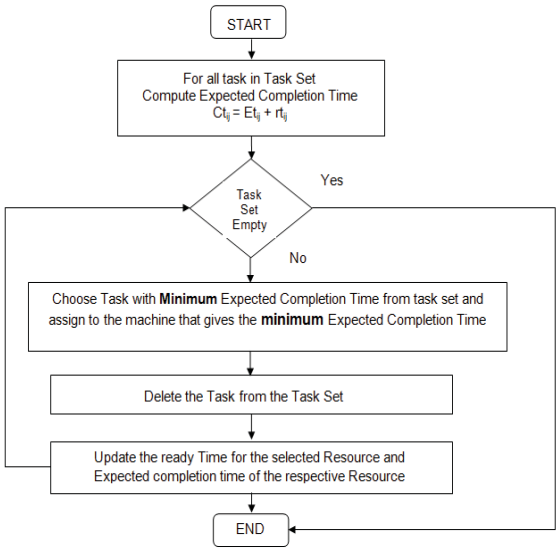


Fig E. Flow chart of MinMin Algorithm

2.3.2. Load balancing based Round Robin

It is a simple algorithm which distributes the incoming requests or traffic in multiple servers circular manner (lets take example of a cricket stadium for better understanding). Each server or resource is selected in turn and receives an equal share of the incoming traffic. Round Robin load balancing can be implemented with or without session persistence and can be effective in evenly distributing traffic among servers or resources, but may not be as efficient in handling variable workloads or optimizing resource utilization. It is a popular load balancing technique for its simplicity and ease of implementation.

In [3], authors did something new in the traditional algorithm rather than taking current response time method they started using modified optimize response time method The method determines the scheduling process after calculating the Reaction Time and Waiting Time for each process. Although it can decrease Reaction Time, but unfortunately this method does not solve the time quantum problem in RR. Which makes it less suitable for dynamic cloud environments.

In Kaur and Yadav [2], a Genetic Algorithm (GA)-based improvement to RR is presented (2019). It seeks to offer effective load balancing so that Data Centers can function more effectively. The method distributes by looking over the hash map, which effectively contains all VMS requests in order to address the problems with RR.

Researchers have improved quality of service in cloud applications by taking out the workload problem into account in the study Issawi et al. [4]. Due to the issue with load balancing that results from an unexpected rise in cloud service customers, cloud computing should take into account both scenarios. As a result, an algorithm known as Adaptive Load Balancing is proposed to efficiently distribute tasks received by VMS by alternating between random task scheduling rules (if workload is normal) and RR task scheduling policies (if workload is bursty) under a strong (bursty) workload. Adaptive LB results minimise reaction time, but when RR is used, a static quantum is enforced, lengthening the waiting period.

2.3.3 Weighted Round Robin

This section outlines the idea behind the researchers' suggested Weighted Round Robin algorithm.

This method, while effective for estimating waiting times, does not account for shifting task durations to allot for the appropriate VM [5]. The algorithm's 694.82 ms Reaction Time [6] is a bit on the high side. Task-based load balancing based on the RR approach is described in [7]. The proposed method, an Enhanced RR, keeps the most recent item provided by a user base in a hash map in order to reduce overall Reaction Time in cloud applications. To reduce waiting and reaction times, the productive load balancing strategy, which combines Max-min and Weighted Round Robin, is suggested by [8]. It selects the activities that demand the most processing time using Max-Min, and then employs weighted RR. both time and reaction time. It selects the activities that demand the most processing time using Max-Min, and then employs weighted RR.

2.3.4 Cloud Partition Based Load Balancing: [9]

As name suggests Partition it works like divide and conquer rule.it makes one principal controller which will control all the partitions, it mainly divides the cloud into four partitions. which results in the optimum scheduling and load balancing. There are two algorithms in it: The optimal partition for work distribution is chosen using the Partition Based Load Balancing Algorithm, and the best refresh interval for the system's load-updating process is chosen using the Determination of Refresh Period approach. Instead of defining any division rules, the technique aims to speed up execution.

2.3.5 Weighted Active Monitoring Load Balancing [10]

This technique functions well in heterogeneous situations where the speed, bandwidth, and number of CPUs of each VM dictate its weight. So, after we calculated the weight, it looks for the vm which contain highest weight after finding the VM allocation table is updated and the id of the VM which contain highest Weight is sent to the data center. This approach ignores the use of VMS while attempting to accelerate response times in the cloud.

2.3.6 Central Load Balancer (CLB) [11]

The approach that is described takes care of VM priority. When determining the priority, CLB considers VMS variables such as RAM & processor performance. This method is good because it leverages the VMS and connects to all users, however because the VM's priority is fixed, it may not function properly when there are a lot of server changes and urgent priorities need to be taken into account. As a result, there can be congestion in the system.

2.3.7 Dynamic Load Management [12]

This technique makes sure that load balancing is done depending on the current status of the VM in order to reduce response time. The index is then erased after requests have been allocated to the selected VM to guarantee that it is busy. The suggested approach performs better than the ideal VM load balancer, although it might not function in static situations, necessitating further resource optimization.

2.3.8 Dynamic Load Balancing [13]

This algorithm seeks to reduce the MS time and maximize resource utilization. It arranges jobs depending on their length as well as processing performance using a bubble sort algorithm. The tasks are then distributed to VMS using FCFS. When this is finished, LBA is used, and the load on each Virtual Machine is tracked and determined. The strategy works well for resource optimization.

2.3.9 Heuristic-Based Load Balancing Algorithm (HBLBA)[1]

By designing servers depends on the quantity of jobs, proportions, and VM concord to maximise efficiency and select the best VM, it seeks to address the issue of poor work allocation to VMS. The method works well when there are a few tasks in the system, but it

may be ineffective when there are many tasks. Adding more configuration details could also slow down the process.

2.3.10 Qos-Based Cloudlet Allocation Strategy [14]

The suggested approach makes sure that LBA takes place when there is at least one free VM (starvation state) because it functions only with the direct nodes for workload balancing. As a result, there are less need of travelling form VM to VM, which eventually save the resources. However, the method is only suitable for independent tasks.

2.3.11 Cluster Based Load Balancing [15]

If a VM is not available to handle the task, it can be shared across other VMS to speed up response times. Although k-means clustering is used in this approach, only Variables and not tasks are clustered. The cluster indicates the VMS's minimum and maximum capacities. The approach makes the list more dynamic by lowering the overhead associated with scanning the complete list in VMS.

3. Discussion

This part discusses a thorough analysis of the algorithms, tables that summaries the studied algorithms, the execution tools that are now available based on the review, and ideas for further research.

3.1 Synopsis of the reviewed algorithms

After examining the algorithms outlined in the literature review sections, the authors sorted the material based on pertinent research gaps. To call attention to prospective future research by other researchers for more breakthroughs, the research gap has been emphasized. The comparison study is then laid out in tables with the names of the algorithms (as proposed by earlier academics), their advantages and disadvantages, and finally name of the author and year of their publication. Response time, fault tolerance, quality of service, priority, and others (such as Makespan, Waiting Time, etc.) are the four characteristics used to classify existing algorithms.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

3.1.1 Fault Tolerance

An effective LBA must have the ability to tolerate faults in any number of nodes in order to work effectively and maintain workload balance. Failures in cloud computing networks can happen for a variety of causes, including system failure, a misconfigured load balancer. There are several efficient ways to deal with errors in cloud environments, like using the replication idea.

No.	Algorithm	Advantages	Disadvantages
1.	Modified Optimize Response Time	Reduce reaction time by using the mapping and sorting technique can be used for social networking sites like Facebook and other ones; efficiently use VM resources	Static weight is employed, and only Memory, bandwidth, and speed are taken into account when calculating weight.
2.	Improved RR	Decrease machine costs, minimise reaction time, avoid overloading, and maximise services.	Tasks with varied configurations are not observed by static algorithms.

3.1.2 Quality of Service

Another crucial parameter for load balancing is QoS. Demand for cloud services is constantly rising, and CSPs have a duty to maintain excellent quality for happier customers. QoS may be affected by a variety of factors, including throughput, latency, availability, and dependability. Moreover, SLA-related parameters such as the Deadline can guarantee excellent QoS by ensuring that user requests are fulfilled in a timely manner. According to the literature, this statistic still needs further investigation because certain algorithms still cause latency problems and don't prioritise user requests.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

No.	Algorithm	Advantages	Disadvantages
1.	Genetic Algorithm (GA)	Minimal amount of time needed to complete user requests.	When search space is expanded, efficiency decreases; job priority is not taken into account.
2.	Central Load Balancer	Better load allocation in a large-scale setting; VMS priority is taken into account.	The fixed priority rating of VMS does not take urgent priorities into account and could result in congestion.
3.	Dynamic Load Balancing	Decrease Makepsan while distributing the workload among resources that are used more frequently.	It works on first come first serve types

3.1.3 Response Time

In cloud computing, CSPs must make sure that the client's needs are met, for instance, if an application user submits a job, it must be finished quickly. Reaction Time is therefore a crucial measure in the load balancing process. According to an analysis of algorithms, it is still need to do additional research in order to further lower response times. The examined algorithms in the table below illustrate this finding.

No.	Algorithm	Advantages	Disadvantages
1.	Dynamic Load Management	Reduces reaction time, using a dynamic virtual machine set. Request allocation occurs faster.	Requests continue to be assigned to overloaded VMs; static environments are not acceptable.

3.1.4 Other

Additional unresolved problems with load balancing, such as Waiting Time, Makespan, and Processing Cost, have been identified in the literature and are described in the algorithms in Table below.

No.	Algorithm	Advantages	Disadvantages
1.	Adaptive LB (RR + Random)	reduces work completion times and makes greater use of available resources.	limitations of the RR algorithm, which uses static quantum and extends waiting time. results in a long time for data processing
2.	Cloud Partition Based Load Balancing	increases system throughput, effectively uses resources, and makes better use of the refresh period policy for job scheduling	It lacks predefined partitioning criteria and is ineffective in a variety of environments.
3.	Weighted Active Monitoring Load balancing	It works in heterogenous environment	It ignores the use of VMS
4.	Heuristic-based Load Balancing	reduces work completion times and makes greater use of available resources.	Not suited for a greater number of tasks, processing time may be slowed down by dependence on other information, such as host settings.
5.	Cluster-Based Load Balancing	maximises resource utilisation; satisfies consumer demand; May be used by the client side	No performance evaluation or evaluation against alternative algorithms

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

In order to identify a new and accurate research gap and move forward with future research in the field of LBA, this study exclusively reviewed literature from the last eight years. Load balancing is one of the crucial things to consider. DLB or NLB should be used in a suitable load balancing algorithm. The research presented several load balancing methods in this work, together with the environment that underlies them. This study demonstrates how recent research is attempting to apply NLB algorithms, which are looed as a metaheuristic strategy, what is metaheuristic strategies, metaheuristic strategies which are used to take forward the solving of optimization problems which includes task scheduling, load balancing, and neural network. Single-objective load balancing algorithms are those that only consider one performance parameter, as opposed to multi-objective algorithms that consider many performance parameters. In this literature study, multi-objective algorithms predominate.

With a focus on lowering Reaction Time in cloud applications, this paper seeks to assist upcoming academics working in the subject of load balancing. One of the difficulties with distributed systems. It displays the overall time taken to reply to a request made by a user or client. It is observed in LBA that which LAB contains dynamic environments needs quicker response time than other strategies.

Secondly it looks for Resource utilization which establishes the similarity in cloud data centres. With a focus on lowering Reaction Time in cloud applications, this paper seeks to assist upcoming academics working in the subject of load balancing. Thirdly the most crucial and difficult task is balancing the Response time in distributed system. It displays the overall time taken to reply to a request made by a user or client. It is observed that LBA that are dynamic in nature require faster response time than others.

No.	**Algorithm**	**Environment**			**Performance Metrics**								
		Static load Balancing (SLB)	Dynamic load Balancing (DLB)	Nature inspired load Balancing (NLB)	Ru	S	TP	RT	MS	AO	FT	MT	SLA

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

1.	Modified Optimize Response Time [3]	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗			
2.	Genetic Algorithm (GA) [2]	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✓			
3.	Improved RR [7]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗			
4.	Adaptive LB (RR + Random) [4]	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗			
5.	WMaxMin [8]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗		
6.	Cloud Partition Based Load Balancing [9]	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗		
7.	Weighted Active Monitoring Load balancing [10]	✗	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗		
8.	Central Load Balancer	✗	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗		

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

	[11]														
9.	Dynamic Load Management [12]	✗	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗		
10.	Dynamic Load Balancing [13]	✗	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗		
11.	Heursitic - Based Load Balancing [1]	✗	✓	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗		
12.	Cluster-Based Load Balancing [15]	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗		

When a feature is present in a review paper, it is indicated by the symbol *(✓); while it is absent, it is indicated by the symbol *(X). Ru: Resource Utilization, RT: Response Time, TP: Throughput, SLA: Service Level Agreement, MS: Makespan, S: Scalability, MT: Migration Time, FT: Fault Tolerance, AO: Associated Overhead.

3.2 Present execution Tools

Tools creates a virtual environment to assess & validate detailed investigations for superior & successful application solution. A technique employed in science is creating a model or a real-time system. The accompanying costs of computer facilities for performance evaluation and modelling the research solution are thus no longer necessary.

To improve cloud computing performance and model load balancing, researchers employed the CloudSim tool. CloudAnalyst, a feature of the CloudSim programme, is second choice of researchers. CloudAnalyst is superior to CloudSim for testing and simulating performance aspects including VM migration and response time. Additionally, it extracts the findings in PDF or XML format, which clarifies and details the task explanation.

By using these implementation tools researchers stated their research according to the performance metrics.

4. Suggestion

The literature suggests that there are still unresolved research difficulties that need to be handled in the future. To further optimization of the resources. Researchers should apply the study's suggestions for bettering load balancing algorithms in their future work while studying cloud computing. Here are a few examples:

1 Failures in the algorithm may occur for a variety of reasons, including an unexpected increase in the number of nodes in the cloud data center, a high priority task that is awaiting execution, an unexpected shift in the workload, and VMS configuration settings.

2 The Particular Round Robin algorithm still belong to static algorithms that cause longer waiting periods. Where min-min algorithm belongs to dynamic algorithm which is slightly better than static algorithms. Researchers can investigate how intelligent algorithms, such nature-inspired algorithms that can handle challenging optimization issues, can be utilized to emulate the cloud environment to address this issue.

The above-mentioned details bring to a close a fresh research hole that specialists in this field may take into account in order to further optimize and improve the execution of Cloud Computing applications. As load balancing aids in cloud optimization, energy-conscious job distribution benefits a variety of cloud applications, particularly those that are crucial for the medical related fields.

5. Conclusion-

By now we understand the importance of load balancing in cloud environment, as of now Cloud computing becoming the primary source of data storage and other services for big companies as well as for common man. Load balancing enhances the workload management, utilizing the available resources at its peak, that can return reduced the system's overall reaction time. while dealing with issues connected to a load balancing, similar as job scheduling, migration, resource utilisation, and others, numerous solutions and techniques have been proposed. The most important problems with load balancing were investigated through a comparison of the strategies put out by researchers during the last 6 years. Despite the numerous techniques that have been used, several issues with the cloud environment still exist, similar as the migration of VMS and problems with failure tolerance.

Researchers have a wide range of options to create clever and effective load balancing algorithms for cloud systems thanks to this review paper. Given that it contains detailed description of ongoing used & can be available load balancing strategies, this study can be useful for researchers to discover research challenges connected to load balancing, particularly to further reduce response time and prevent server failures.

Reference

1 Adhikari, M., Amgoth, T., 2018. Heuristic-based load-balancing algorithm for laaS cloud. https://www.sciencedirect.com/science/article/pii/S131915782100046X

2 Kaurav, N.S., Yadav, P., 2019. A genetic algorithm-based load balancing approach for resource optimization for cloud computing environment. https://ijics.com/-gallery/26-mar-965.pdf

3 Tailong, V., Dimri, V., 2016. Load balancing in cloud computing using modified optimize response time. https://www.sciencedirect.com/science/article/-pii/-S131915782100046X

4 Issawi, S.F., Al Halees, A., Radi, M., 2015. An efficient adaptive load balancing algorithm for cloud computing under bursty workloads. https://etasr.com/index.php-/ETASR/article/-view/554

5 Mayur, S., Chaudhary, N., 2019. Enhanced weighted round robin load balancing algorithm in cloud computing https://www.ijitee.org/wp-content/uploads/papers/-v8i9S2/I10300789S219.pdf

6 James, J., Verma, D.B., 2012. Efficient Vm load balancing algorithm for a cloud computing environment. https://www.researchgate.net/publication/265266981-_EFFICIENT_VM_LOAD_BALANCING_ALGORITHM_FOR_A_CLOUD_COMPUTING_ENVIRONMENT

7 Pasha N., Agarwal, A., Rastogi, R. , 2014. Round Robin Approach for VM Load balancing algorithm in cloud computing environment. https://www.semanticscholar.-org/paper/Round-Robin-Approach-for-VM-Load-Balancing-in-Cloud-Pasha-Agarwal/822dc4e2755ccfe93a1f5b9cec9c4a7470a94d51

8 Khatavkar, B., Boopathy, P., 2017. Efficient WMaxMin static algorithm for load balancing in cloud computation. https://www.researchgate.net/publication/-322350309_Efficient_WMaxMin_static_algorithm_for_load_balancing_in_cloud_computation

9 Chaturvedi, M., Agrawal, P.D., 2017. Optimal load balancing in cloud computing by efficient utilization of virtual machines. https://www.sciencedirect.com/-science/article/-pii/S131915782100046X

10 Singh, A.N., Prakash, S., 2018. Wamlb: weighted active monitoring load balancing in cloud computing. https://www.researchgate.net/publication/320213402_WAMLB_-Weighted_-Active_Monitoring_Load_Balancing_in_Cloud_Computing

11 Soni, G., Kalra, M., 2014. A novel approach for load balancing in cloud data center https://www.researchgate.net/publication/271482427_A_novel_approach_for_load_balancing_in_cloud_data_center

12 Panwar, R., Mallick, B., 2016. Load Balancing in Cloud Computing Using Dynamic Load Management Algorithm https://doi.org/10.1109/1CGC10T.2015.7380567

13 Kumar, M., Sharma, S.C., 2017. Dynamic load balancing algorithm for balancing the workload among virtual machine in cloud computing. https://www.sciencedirect-.com/science/article/pii/S1877050917319695

14 Banerjee, S., Adhikari, M., Kar, S., Biswas, U., 2015. Development and Analysis of a new cloudlet allocation strategy for QoS improvement in cloud https://doi.org/-10.1007/s13369-015-1626-9.

15 Kamboj, S., Ghumman, M.N.S., 2016. An implementation of load balancing algorithm in cloud environment. https://www.researchgate.net/publication/-324987491_AN_IMPLEMENTATION_OF_LOAD_BALANCING_ALGORITHM_IN_CLOUD_ENVIRONMENT