Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Comprehensive Review Based on Load Balancing in Cloud Computing

Malavika N
Dept.Computer Science and Engineering
Vimal Jyothi Engineering College, Chemperi

Abstract - **Cloud Computing stands as a resilient model that empowers users and organizations to procure services tailored to their specific needs. Within this model, an array of services, including storage, deployment platforms, and seamless access to web services, is offered. However, the challenge of Load Balancing in the cloud complicates the maintenance of application performance in alignment with Quality of Service (QoS) metrics and adherence to Service Level Agreements (SLA) specified by cloud providers for enterprises. Achieving an equitable distribution of workload among servers poses a continual challenge for cloud providers. To ensure optimal resource utilization of Virtual Machines (VMs) and enhance user satisfaction, an efficient Load Balancing (LB) technique is imperative. This paper presents a comprehensive examination of diverse Load Balancing techniques within static, dynamic, and nature-inspired cloud environments, specifically focusing on addressing Data Center Response Time and overall performance. The analysis includes an in-depth evaluation of these algorithms, pinpointing their strengths and limitations, while also identifying potential avenues for future research. Additionally, the research incorporates graphical representations of the operational flow of reviewed algorithms. Moreover, the paper introduces a fault- tolerant framework and explores existing frameworks in recent literature, providing a holistic view of Load Balancing strategies in the evolving landscape of cloud computing.**

I. INTRODUCTION

Cloud Computing stands out as a leading technology, providing both private and public services that facilitate seamless access to data, programs, and files over the internet (cloud). Scalable storage solutions are offered online, negating the need for local storage on user devices such as computers or phones. Originally proposed by Prof. Ramnath Chellappa in 1997, this technology has evolved to offer dynamic services, including cost-effective and scalable alternatives, as highlighted by var- ious studies (Agarwal and Srivastava, 2017; Nazir, 2012; Ab- dalla and Varol, 2019). Businesses globally benefit from Cloud Computing's aim to reduce hardware costs, employing a Pay- Per-Use model where clients can purchase services according to their specific needs, often through subscription models. Notable technology companies like Google, Microsoft, and IBM widely adopt this model, particularly in the Software as a Service (SaaS) delivery model (Lowe and Galhotra, 2018). Fig. 1 below provides a summary of Cloud Computing, illustrating the collaboration of various entities within the cloud environment. Cloud auditors act as overseers, ensur- ing high-quality and integrity in services provided by Cloud Service Providers (CSPs), while cloud carriers ensure stable connections to transport services to clients (cloud users). In the private cloud, the Data Center is within the organization's network, while in the public cloud, it relies on the internet and depends on CSPs. Hybrid clouds can have Data Centers in both environments.

In a typical Cloud Computing environment, two main components are identified: the frontend and the backend. The frontend, accessible to users over the Internet, contrasts with the backend, which manages cloud service models. User requests are dynamically scheduled, and resources are allocated through virtualization, a technique responsible for handling dynamic resources, load balancing, and efficient resource allocation (Jyoti et al., 2019). Requests from users, transmitted via the internet, are stored in Virtual Machines (VMs), and CSPs, in each delivery model, must uphold Quality of Service (QoS) by ensuring timely execution of user requests (Adhikari and Amgoth, 2018). The allocation of user tasks to VMs depends on a scheduling policy (Data Broker) that strives for a balanced workload across machines and servers. The design and implementation of a dynamic load balancer contribute to efficient scheduling and resource utilization.

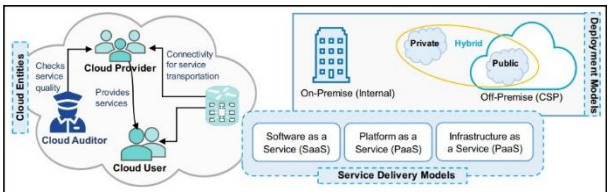

Fig. 1. Overview of cloud computing.

The study makes a significant contribution through a comprehensive survey of 58 existing load balancing algorithms, providing insights into their respective strengths and issues. Additionally,the analysis is extended by incorporating findings from existing review papers, presenting flowcharts depicting the operational flow of these algorithms, and compiling ex- perimental results based on specific metrics, such as Response Time and Processing Time. Furthermore, a novel framework aimed at addressing the critical fault tolerance issue in load balancing algorithms are introduced. This framework not only enhances migration techniques but also mitigates node failures. While researchers have previously touched upon this issue, the predominant focus

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

has often been on the utilization of a single load balancer. In contrast, the proposed fault-tolerant model tackles failures by employing dual Load Balancers and leveraging machine learning tools to predict potential issues in the active Load Balancer. Although some attention has been given to this concern in previous research, the emphasis has typically been on a singular load balancer.

II. LITERATURE SURVEY

A. Load balancing based Throttled algorithm

The Throttled Algorithm (TA) operates as a dynamic load balancing algorithm, tasked with identifying a suitable Virtual Machine to execute tasks upon receiving a client request. TA manages an index table,known as the allocation table. When an available VM with sufficient capacity is found, TA allocates the task to it. If no suitable VM is available, TA returns - 1, and the request is queued for expedited processing. While TA outperforms the Round Robin algorithm, it falls short in addressing more advanced load balancing requirements, such as Processing Time. Similar to the traditional TA algorithm, the Modified Throttled Algorithm maintains an index table of all VMs and their states. If the VM at the first index is available, the request is assigned, and -1 is returned to the Data Center. Subsequent VMs are selected in order. This differs from traditional TA, where the VM at the first index is consistently chosen for every request. Researchers introduced a priority-based approach, the Priority Modified Throttled Algorithm (PMTA). PMTA focuses on task allocation by utilizing a switching queue to prioritize high-priority tasks over low-priority ones, aiming to achieve equal workload distribution across multiple VMs. The Throttled with Multiple Allocation algorithm addresses equal workload distribution by maintaining two tables of VMs with states: available and busy. TMA facilitates easier detection of VM availability. The algorithm slightly reduces Response Time indicating the need for further optimization to enhance overall performance

B. Load balancing based Equally Spread current execution

The Throttled Algorithm has been proposed as a solution to address the load balancing challenges in cloud computing. Another dynamic algorithm in this domain is Equally Spread Current Execution (ESCE), which prioritizes job size and randomly distributes workload to Virtual Machines with lighter loads, referred to as the Spread Spectrum technique. ESCE employs a queue to manage requests and distributes loads to VMs, particularly targeting overloaded ones. The drawback of ESCE lies in potential overhead during the updating of the index table, resulting from communication between the Data Center controller and the load balancer. Researchers have explored a hybrid approach, combining Equally Spread Concurrent Execution (ESCE) with TA, aiming to reduce Response Time in Cloud Computing (CC). The Hybrid Load Balancing approach utilizes a HashMap list to store user requests and scans through it to find an available VM. Unlike TA, which returns -1 if VMs are busy, ESCE is employed to locate a machine with minimal load to assign the task. Similar approaches have been proposed by different authors, such as Enhanced Load Balancing, which optimizes resources by replacing the queue with a create host function to reduce waiting times. However, both these approaches do not address VM migration in the event of faults.

Load Balancing Hybrid Model (LBHM) is introduced as another hybrid approach using TA and ESCE, aiming to reduce waiting times when the task number increases. LBHM sets a threshold limit for each VM based on its capacity, but it struggles in scenarios involving node failures. Further variations of the hybrid approach include a Hybrid Approach, a Hybrid Virtual Machine Load Balancin. All of these combine TA and ESCE to achieve better workload distribution and response time, with considerations for fault tolerance, resource consumption, and optimal performance.

C. Load balancing based round Robin

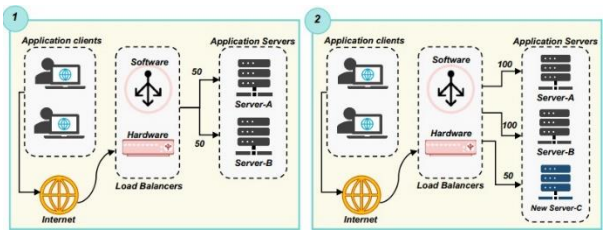

Fig. 2. Drawback of RoundRobin

The Round Robin Algorithm, operates in a cyclic and ordered manner, assigning each process a fixed time slot without any priority. Despite its simplicity, a common issue in load balancing arises from the lack of saving and updating the VM's allocation state after user requests. The static nature of RR becomes evident in a compiled figure where additional user requests result in an uneven distribution of load among servers.

A Modified Optimize Response Time algorithm is proposed to alter the existing Response Time service broker policy in CloudAnalyst. While effective in reducing Response Time, it does not address the time quantum problem in RR, making it less suitable for dynamic cloud environments. By scanning through a hash map containing VM information, the algorithm allocates tasks efficiently, resorting to GA for selecting the best-fitted tasks when necessary. Despite significantly reduc- ing server Response Time, the complexity of GA increases with larger search spaces. This algorithm dynamically adapts between RR and Random scheduling policies to ensure equal distribution of tasks to VMs under varying workload condi- tions. While Adaptive LB reduces Response Time, the use of RR with a static quantum results in increased waiting time.

The Improved RR approach incorporates two processors: a small processor for calculating time slices and a main processor arranging processes based on burst time priority. While aiming to reduce Response Time, there is a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

lack of performance testing in the literature.

D. Load balancing based Ant Colony Optimization Algorithm

Ant Colony Optimization (ACO) is inspired by the foraging behavior of ants during their quest for food. As ants roam randomly in search of food, they deposit pheromones upon their return, marking the shortest path from the nest to the food source. While this approach is effective for resource optimization, it tends to result in slow Response Time and performance issues. T

In an effort to achieve effective scheduling and equal distribution among servers, a hybrid algorithm that combines ACO with the priority ABC algorithm, forming the Hybridized ACO Priority-based Bee Colony. This hybrid approach assigns priority to tasks based on the shortest job criteria, akin to the communication of the shortest distance among nodes by ants and bees. While the proposed algorithm using CloudAnalyst reduces response time and cost, there is room for further enhancement to minimize Data Center Processing Time. The algorithm allocates requests to Virtual Machines (VMs) using the First-Come-First-Serve (FCFS) algorithm and employs an index table for storage. Ants traverse randomly to identify the optimal VM for allocation. While designed to enhance Quality of Service (QoS) for customer jobs, the algorithm assumes equal priority for all jobs, which may lead to potential failures in the system.

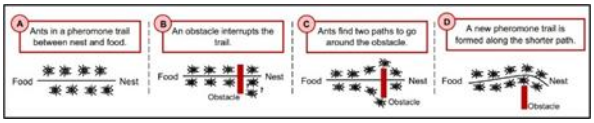

Fig. 3. Drawback of RoundRobin

E. Load balancing based Particle Swarm algorithm

Particle Swarm Optimization (PSO) emulates the natural gathering behavior observed in populations, such as the forag- ing behavior of birds or ducks gathering to find food. In this context, the population is referred to as the swarm, and the individuals within the swarm are represented as particles, akin to ducks. These particles explore the global space using a set velocity, allowing them to adapt and update their positions. PSO, proven to be highly effective in Neural Network appli- cations, operates with simpler rules compared to the Genetic Algorithm (GA) as it does not involve mutation or crossover operations. PSO seeks the optimal solution through iterative processes, utilizing a fitness function to evaluate the quality of solutions.

A hybrid approach is proposed by combining PSO with the Equally Spread Current Execution Load (ESCEL) algorithm, forming the Hybrid PSO ESCEL. PSO is employed to optimize jobs on the cloud server before task assignment, and then the server uses the ESCEL approach for task allocation. This hybrid approach aims to optimize resources and achieve faster response times. However, it is important to note that there is no experimental evaluation provided to substantiate the effectiveness of this proposed method.

III. CONCLUSION

Load balancing plays a crucial role in the realm of Cloud Computing, aiming to enhance workload distribution and optimize resource utilization, ultimately leading to a reduction in the overall system response time. Numerous approaches and algorithms have been proposed to address various load balanc- ing issues, including task scheduling, migration, and resource allocation. This study explores several strategies tackling the significant challenge of load balancing in Cloud computing. Through a comparative analysis of algorithms proposed by researchers over the past six years, the study highlights the diverse methodologies employed.

While a variety of approaches have been suggested, some challenges persist in the cloud environment, such as the migration of Virtual Machines (VMs) and inadequately ad- dressed fault tolerance issues. This review paper underscores the existing gaps in addressing these challenges.

By providing a comprehensive overview of load balanc- ing techniques, this review paper offers a valuable resource for researchers seeking to develop intelligent and efficient algorithms tailored for cloud environments. It serves as a reference for identifying research problems related to load balancing, with a particular emphasis on further reducing response time and preventing server failures. The summary of existing load balancing techniques encapsulated in this study provides a foundation for future research endeavors in this critical domain.

REFERENCES

[1] ”Load Balancing in Cloud Computing: A big Picture” 1Sambit Kumar Mishra, 1Bibhudatta Sahoo, 2Priti Paramita Parida , 2018
[2] ”Load balancing techniques in cloud computing environment: A review”
, Dalia Abdulkareem Shafiq, N.Z. Jhanjhi, Azween Abdullah , 2021
[3] ”Load balancing in cloud computing a load aware matrix ap- proach”,Muhammad Sami Ullah, 2017
[4] ”Load balancing in Cloud Computing: Issues and Challenges” ,
K.Balaji1, P.Sai Kiran 2, M.Sunil Kumar 3 , 2021
[5] ”Load Balancing in Cloud Computing with Enhanced Genetic Algo- rithm” Kalpana, Manjula Shanbhog 2019