Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS 2023)
IEEE Xplore Part Number : CFP23CB2-ART ; ISBN : 979-8-3503-2579-9

2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) | 979-8-3503-2579-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAISS58487.2023.10250629

Contemporary Load Balancing Mechanisms: A Comprehensive Review

[1] Kesava Vamsi Krishna V.
Dept. of Physics,
Malla Reddy Engineering College,
Secunderabad, India
mrecphysics@gmail.com

[2] M.Sravanthi
Dept. of ECE
Malla Reddy Institute of Engineering and Technology, Hyderabad, India
sravanthi.engg@gmail.com

[3] D.V.Acharyulu
Dept. of Pysics
Malla Reddy Engineering College,
Secunderabad, India
sahithividyaranya@gmail.com

[4] G.Vijaya Nirmala
Dept. of ECE
CVR College of Engineering,
Hyderabad, India
g.vijayanirmala@cvr.ac.in

[5] J.Jaganpradeep
Dept. of ECE
SSM College of Engineering
Komarapalayam, Tamilnadu, India
jgnprdp@gmail.com

[6] G.Revathi
Dept. of ECE
SSM College of Engineering
Komarapalayam, Tamilnadu, India
revathirenu2007@gmail.com

Abstract— **Load Balancing (LB) mechanism in cloud computing distributes workloads across multiple computing resources such as, Servers and Virtual Machines (VM). The ultimate goal of LB mechanism is to provide the optimum usage of resources in cloud and ensure none of the resource is overburdened even in high network traffic. The significant feature of LB mechanism encompasses uniform workload distribution thereby improving the performance of the system and ensures high availability of servers; thus, guarantees fault tolerance. Moreover, LB scales up or down the resources as and when required based on the spike in the network traffic. The proposed study provides an insight on LB mechanism, its types and the techniques adapted in LB to attain optimum efficiency or to reach the goals of LB mechanism. Further, this study focuses on the LB mechanism that optimizes the resource usage to bring down the operational costs. The load balancers are categorized as static and dynamic. The choice of LB depends upon the parameters like, resource utilization, workload allocation, priority of workloads as some works relies on other workloads, compilation time, execution time, energy consumption and throughput.**

Keywords— Load balancing, Optimization, Throughput, Fault tolerance, Quality-of-service, Legal service agreement.

I. INTRODUCTION

Load Balancing (LB) is a technique to distribute the network traffic equally among server farm or pool of servers to deliver optimal network performance, reliability and capacity ensuring minimum latency [1].

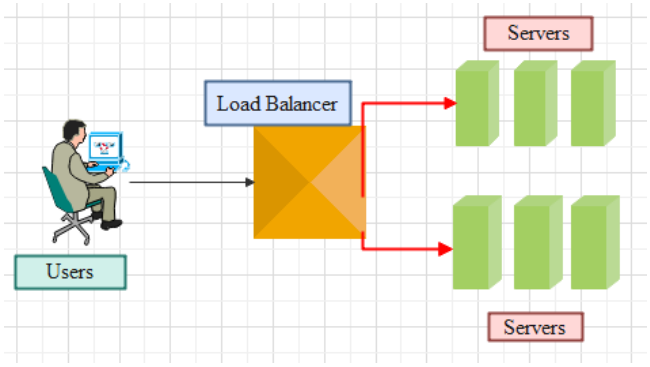


Fig. 1. System Design

A. LB Workflow

Load balancers handle user requests for information or service. Usually load balancers work between the servers with the request and the internet. The progress of load balancing is as follows: (1) Upon incoming request, load balancer determines the available server online (2) routes the request to the determined server. Based upon the spikes in traffic, the load balancers can add servers and reduce servers. Figures 1 and 2 depict the LB System design and representation of LB processing across distinct servers.

Technically, the working of LB mechanism is illustrated in Figure 3.

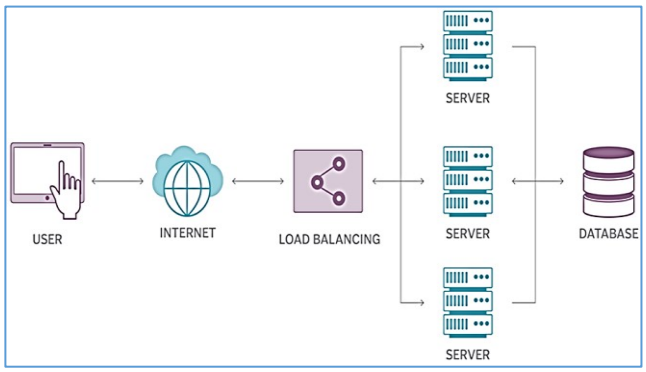


Fig. 2. Representation of LB Processing across distinct Servers (Source: Courtesy: [2])

B. LB: Types

Load balancing uses either a virtual or physical component to identify an ideal server from the server farm for the client request thus avoiding heavy traffic for a specific server.

Load balancing is usually supported either in Layer 7 or Layer 4 in the OSI (Open Systems Interconnection). Layer 7, a common load balancer handles routing decisions like Hyper Text Transfer Protocol header information, URLs and actual content. Layer 4 load balancers which play pivotal role in edge deployments, transmit IP addresses and TCP port numbers. Figure 4 depicts the LB types.

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on August 11,2026 at 09:50:03 UTC from IEEE Xplore. Restrictions apply.

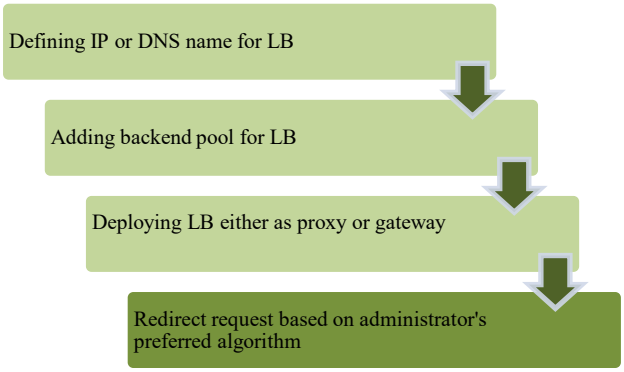


Fig. 3. LB: Working Principle

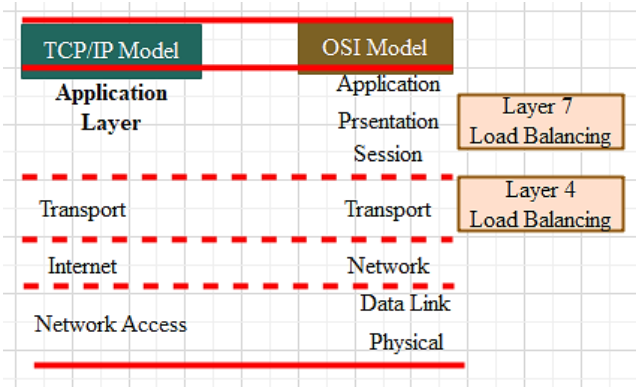


Fig. 4. Types of LB

C. L4 Load Balancers

L4 LBs acts based on data available in network and transport layer protocols. These Network Address Translators (NAT) LBs share the load to different servers. Session endurance can be attained at IP address level. L4 LBs work in two modes namely, DSR mode and NAT mode. Figures 5 and 6 illustrates DSR and NAT mode, respectively.

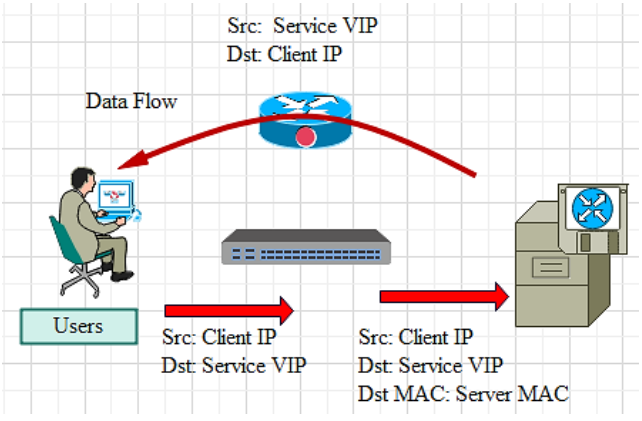


Fig. 5. DSR Mode (Source: Courtesy: [15])

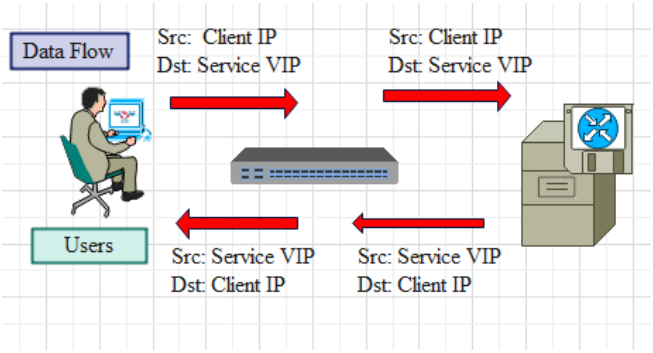


Fig. 6. NAT Mode

II. Retrospect of Contemporary LB Mechanisms

The following section depicts the retrospect of contemporary LB Mechanisms.

Oludayo A. Oduwole and team members [3] conducted a detailed investigation on peer-review papers available in Google Scholar on cloud computing Load Balancing techniques. Among 201 papers searched 39 papers were found suitable for the investigation. The investigation involved detailed study on Load Balancing (LB) Algorithms such as Pre-emptive approach, Responsive LB, Static and Dynamic LBs, Centralized approach, Distributive Technique, Hierarchical LB.A detailed comparison of afore mentioned techniques on the basis of efficiency, scalability and fault tolerance was carried out. The exhaustive investigation revealed the flaw in optimizing processing cost and cloud architecture. To overcome, a central distributive framework with maximum throughput was present by the authors.

Ashawa, Douglas, Osamor and Jackie [4] improved cloud efficiency by means of optimized resource allocation using Long Short Term Memory (LSTM) algorithm. Moreover the results were integrated with dynamic routing algorithm focussing on handling cloud data traffic. The efficiency levels of LSTM were compared against Monte Carlo Tree Search method. As traffic patterns might shift rapidly, consistency maintenance throughout the simulation was difficult in MCTS. In LSTM, the issue was handled successful and Service Legal Agreement (SLA) was attained. When LSTM was compared against other LBs the accuracy rate improved by 10-15% with the error rate of 9.5-10.2% thus implying LSTM a better predictive approach in improving the network usage.

Shinde and Ranbhise [5] analysed the techniques of Load Balancing in Cloud and ascertained improved weighted Round Robin LB algorithm as an efficient LB algorithm. The authors briefed Cloud under three major components namely Cloud Service Models, Deployment Model and its components in addition to the Virtualization techniques- Full and Partial Virtualization. A LB algorithm commits to Transfer, Selection, Location, Information and Load Estimation Policies to attain maximum resource utilisation in minimum response time with higher throughput. The authors spotlighted the LB algorithms such as Min-Min, Round Robin, Weighted Round Robin, Task Scheduling, Opportunistic, Randomized, Max-Min, First come First Serve, Shortest Response time First. The scheduling technique plays pivotal role in Load Balancing for optimising the resource usage thus avoiding over and under usage of servers

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on August 11,2026 at 09:50:03 UTC from IEEE Xplore. Restrictions apply.

in cloud. To wrap up, the authors justified improved Weighted Round Robin allocated the jobs to Virtual Machines suitably as it met the goals of an ideal LB in terms of cost, Scalability, flexibility and priority.

Anusha, Bindu and Nagamani [6] suggested Round Robin algorithm as efficient technique for LB. The function SA (Service Adaptor) determines acceptance or denial of service, once the service reaches the cloud. Upon acceptance of the service, SA checks for availability of VM (Virtual Machine). If VM is free, then Round Robin algorithm performs scheduling work. The ultimate goal of LB to increase the performance of cloud by utilising the resources to maximum extent was attained by Round Robin. Subsequently, the number of job rejection was reduced.

Sharma and Joshi [7] devised a hybrid algorithm using WRR and Scheduling algorithms to attain optimized LB with minimum energy consumption thereby accomplishing Green Cloud Environment. The authors initially determine the number of services, simulation time, workflow determination, determination of VMs for services, computation of execution time for every service, queuing up the services and designate VMs for processing. Subsequently implement Weighted Round Robin (WRR) and MCT (Minimum Compilation Time) hybrid algorithm. In other words, priority based execution of services. Finally, in the cloud environment, the performance evaluation of completed services, incomplete services, minimum time consumed by the completed services, minimum compilation time of the completed services in the proposed algorithm was compared with MCT and MET techniques. The hybrid WRR and scheduling algorithm reduced energy consumption and throughput.

Kanbar and Faraj [8] discussed modern load balancing techniques and their impact on cloud computing. The authors analysed the effectiveness of Round Robin, Least Connection, Weighted Round Robin, IP hash, shortest job first, least connection, and dynamic load balancing and their impact in cloud. Load balancing, an effective technique in cloud computing improve the performance of cloud services by reducing response time, increasing throughput, and minimizing downtime. To provide customer satisfaction, selection of appropriate LB on the basis of cloud infrastructure and workload rely on cloud service provider. In addition, the effectiveness of LB is measured in terms of scalability, fault tolerance, and energy consumption.

Gao and Yu [9] put forth an energy-aware load balancing algorithm for heterogeneous cloud data centers, where servers have varying capacities and energy consumption. The algorithm optimized energy efficiency. The algorithm used a dynamic threshold-based approach to distribute the workload among servers and minimize the number of active servers to reduce energy consumption. The proposed algorithm was evaluated using simulations. On comparison with existing LBs, the proposed algorithm outperformed in terms of energy efficiency. The energy-aware load balancing algorithms reduce the carbon footprint of data centers and contribute to their sustainability. Thus, by optimised energy efficiency and reduced operational costs the algorithm act as a boon to cloud service provider.

Pilavare and Amish [10] proposed a load balancing technique using Genetic Algorithm (GA) where the issue of selecting processors or jobs in random were resolved using Logarithmic Least Square Matrix (LLSM). LLSM computes comparison matrices for the jobs selected in random; priorities for jobs are listed; subsequently priority-based allocation of jobs take place in cloud. The consistent model of GA's random selection of jobs results in starvation and low fitness value of VM thus leading the VM idle. Hence priority-based job input of VM in GA performed amicably than other LBs in simulation experiments and improved QoS for the cloud users.

Alghamdi [11] proposed an LB technique integrating Artificial Neural Networks (ANN) and Binary Particle Swarm Optimization (BPSO) to overcome the failure of conservative LB in handling the dynamic changes in cloud. The integrated LB handles not only the rapid changes of workload in real time also the resource availability. In terms of execution time, response time (51.6%) and makespan (42.1%) the integrated approach outperformed existing techniques enabling it viable for the dynamic and modern cloud computing environments apart from ease of use.

Jing [12] proposed a novel load balancing mechanism for cloud computing that integrates Ant Colony Optimization (ACO) algorithm with the existing load balancing methods. The integrated mechanism addresses the dispute of load imbalance among servers in cloud computing environments resulting in high response time, increased energy consumption, and reduced system efficiency. ACO algorithm mimics the way ants forage for food by laying pheromone trails that other ants can follow. Similarly, ACO balance the load between cloud servers by forming a trail between servers based on the amount of load and the distance between them. The integrated technique outperformed other LBs like RR, exclusive ACO and Throttled in Simulation experiments.

Parul and Abhilasha [13] investigated the limitations of existing LBs and proposed Honey Bee based LB (HBLB)mechanism to improve resource utilization and execution time in cloud. The forage behaviour of honey bee to identify food source and on return its communication to worker bees about the source and its quality was applied LB. In the proposed algorithm, the bees explore the available VMs, gather information about their workloads and processing calibre and finally designate the work to VMs based on the criterion. The algorithm initiates by determining the cloudlet length and categorizing the VMs as overloaded VM, under-loaded VM and balanced VM. Subsequently, cloudlets were assigned to VMs and load on every VM was identified. Monitoring the balance of VM is the highlighting factor in the proposed mechanism. If VM is balanced then further processing happens if at all HBLB reallocates the cloudlets thus ensuring none of the VMs were overloaded. The simulation results of the algorithm before and after applying HBLB for 20 cloudlets on 8 VMs were compared to show the balanced distribution of cloudlet when HBLB was used.

Anjuli Goel and Pooja Kalpesh [14] conducted a detailed review on several LB techniques in Cloud Computing. The authors identified different LBs namely Static, Dynamic and Hybrid algorithms and highlighted their advantages and disadvantages. Static algorithms are suitable where resources and workloads are predetermined while Dynamic algorithms adapt to changes in the workload and resource availability. Hybrid algorithms combine the features of both static and dynamic algorithms to achieve better performance. The algorithms under the aforesaid category such RR, WRR, LL, ACO, PSO, GA were discussed in detail. While selecting load balancing algorithms multiple parameters must be considered including CPU utilization, memory usage, network traffic, and

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on August 11,2026 at 09:50:03 UTC from IEEE Xplore. Restrictions apply.

geographical location, to achieve optimal performance. Moreover, the selected LBs should be scalable and fault-tolerant to handle large-scale cloud environments and unexpected failures.

Jaleel Nazir and his team [15] designed a framework to balance the task loads across multiple regions in cloud computing to enhance LB. The team highlighted and justified geographic location of cloud resource as an important parameter for balanced workload as it contributed more for the overall performance of the framework. The teams designed the frame in 3 phases: task clustering – assembly of clusters that hold similar workloads, region selection – embrace the appropriate region and task scheduling – allots the task to suitable regions' resources [16]. The design used K-Means, Genetic Algorithm and modified GA for the purpose. The 3 phases enable the framework viable for optimum resource utilization, minimum latency and enhanced response time thus meeting the goals of LB. The design was demonstrated in Amazon EC2 cloud infrastructure and compared with other LBs; exhibited reduced response time by up to 41%, outperformed in terms of resource utilization and service availability.

III. Experimental Results and Discussions

Table 1 summarizes the results and significant features of the contemporary LB mechanisms.

TABLE I. Significant features of contemporary LB mechanisms

Ref. No.	Technique	Metric Highlighted
[3]	Review of 39 papers on different LBs	Suggested a central distributive framework with maximum throughput
[4]	• Long Short Term Memory (LSTM) • Monte Carlo Tree Search	• Swift in Cloud data traffic handled well • Accuracy rate improved by 10-15% with the error rate of 9.5-10.2%
[5]	• Improved weighted Round Robin LB • Discussion of other LBs	• Optimum resource usage • Priority, Scalability, Flexibility
[6]	• RR	• Maximum resource utilisation • Reduced job rejection
[7]	• WRR + Scheduling	• Energy consumption • Minimum compilation time • Minimum execution time
[8]	Analysis of RT, LC,WRR, IP hash, shortest job first, least connection, and dynamic load balancing	Effectiveness of LB discussed in terms of scalability, fault tolerance, and energy consumption
[9]	• Threshold-based approach	• Reduced carbon footprint of data center • Optimized energy efficiency • Minimum number of active servers • Minimum operational costs
[10]	• GA	• Priority-based allocation of jobs
	• Logarithmic Least Square Matrix (LLSM)	• Reduced starvation of VM • Low fitness value of VM
[11]	• Artificial Neural Networks (ANN) + • Binary Particle Swarm Optimization (BPSO)	• Dynamic changes in workload • Response time (51.6%) • Makespan (42.1%)
[12]	• Ant Colony Optimization (ACO) algorithm and • Existing LBs	Simulation results depicts • Low response time • Minimum energy consumption and • Increased system efficiency
[13]	• Honey Bee-based LB (HBLB)	Simulation results of the algorithm before and after applying HBLB for 20 cloudlets on 8 VMs in terms of execution time, workload on VMs
[14]	Review of RR, WRR, LL, ACO, PSO, GA	• CPU utilization • Memory usage • Network traffic and • Geographical location
[15]	• K-Means • GA • Modified GA	• Geographic location of cloud resource

IV. Conclusion

The objective of the manuscript was to study the significance of LB mechanism and its impact in the Cloud Computing environment. The observations of this study are as follows. The LB mechanism is a crucial aspect of cloud computing that ensures efficient resource allocation, improved application performance, and optimal resource utilization. Various load balancing techniques and approaches have been reviewed to address the challenges associated with cloud computing environments. These approaches include traditional load balancing techniques, such as round-robin and least connections, as well as advanced techniques, such as, priority-based and optimization algorithms such as, GA, ACO, PSO and HBLB integrated with traditional approaches for load balancing. The traditional load balancing techniques are simple and efficient but do not take into account the dynamic nature of cloud computing environments. Advanced load balancing techniques, on the other hand, offer more intelligent and distributed load balancing decisions based on real-time data analysis and prediction. In recent years, machine learning techniques have been integrated into load balancing algorithms, enabling load balancers to learn from past behavior and adjust their decisions accordingly. Overall, cloud computing load balancing is a complex and dynamic process that requires careful consideration of various factors, such as network topology, server capacity, and workload characteristics. Cloud providers and Information Technology professionals must carefully evaluate the available load balancing techniques and approaches to determine the best solution for their specific needs. In the future, we can expect further advancements in load balancing techniques and approaches, as cloud computing environments continue to evolve and become more complex. These advancements will play a vital role in ensuring the continued growth and success of cloud computing as a key technology for modern businesses and organizations.

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on August 11,2026 at 09:50:03 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] "What Is Load Balancing? How Load Balancing Works | TechTarget," Networking, 03-Jan-2023, https://www.techtarget.com/searchnetwork ing/definition/load-balancing.

[2] "System Design Load Balancing," Medium, https://yangpeng-tech.medium.com, 04-Dec-2022. https://medium.com/must-know-computer-science/system-design-load-balanc ing-1c2e7675fc27.

[3] Oludayo A. Oduwole, Solomon A. Akinboro, Olusegun G. Lala, Michael A. Fayemiwo and Stephen O. Olabiyisi, Cloud Computing Load Balancing Techniques: Retrospect and Recommendations, FUOYE Journal of Engineering and Technology, vol. 7, no. 1, pp. 17-22, 2022.

[4] Moses Ashawa, Oyakhire Douglas, Jude Osamor and Riley Jackie, Improving cloud efficiency through optimized resource allocation technique for load balancing using LSTM machine learning algorithm, Journal of Cloud Computing, vol. 11, no. 87, pp. 1-17, 2022, https://doi.org/10.1186/s13677-022-00362-x.

[5] Krishnajali J. Shinde and Satish Ranbhise, Efficient Techniques for Load Balancing in Cloud, International Journal of Engineering Research & Technology (IJERT), vol. 5, no. 01, pp. 1-6, 2017.

[6] S. K. Anusha, N. R. Bindu Madhuri and N. P. Nagamani, Load Balancing in Cloud Computing using Round Robin Algorithm, International Journal of Engineering Research & Technology (IJERT), pp. NCRTS`14 Conference Proceedings, 124-127, 2014.

[7] Himanshu Sharma and Vijay Kumar Joshi, Load Balancing Optimization for Green Cloud Environment Using Effective Scheduling, Intelligent Systems and Applications in Engineering, vol. 10, no. 1s, pp. 327-334, 2022.

[8] Asan Baker Kanbar, Kamaran Faraj, "Modern Load Balancing Techniques and Their Effects on Cloud Computing", Journal of Hunan University (Natural Sciences, vol. 49, no. 7, pp. 37-43, July 2022.

[9] Yongqiang Gao and Lei Yu, "Energy-aware Load Balancing in Heterogeneous Cloud Data Centers" ISBN 978-1-4503-4834-8/17/01,

[10] Mayur S. Pilavare and Amish Desai "A novel approach towards improving performance of load balancing using Genetic Algorithm in cloud computing," 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, 2015, pp. 1-4, doi: 10.1109/ICIIECS. 2015.7193124.

[11] Mohammed I. Alghamdi, "Optimization of Load Balancing and Task Scheduling in Cloud Computing Environments Using Artificial Neural Networks-Based Binary Particle Swarm Optimization (BPSO)", Sustainability, vol. 14, no. 19, 2022.

[12] Jing He, "Cloud Computing Load Balancing Mechanism Taking into Account Load Balancing Ant Colony Optimization Algorithm", Computational Intelligence and Neuroscience, article ID. 3120883, volume 2022, pp. 1-10, 2022.

[13] Sharma, Parul and Sharma, Abhilasha, Honeybee Inspired Load Balancing Algorithm for Cloud Computing Environment, February 3, 2022. http://dx.doi.org/10.2139/ssrn.4025366

[14] Anjuli Goel and Pooja Kalpesh, Various Load Balancing Techniques in Cloud Computing: A Review, Pramana Research Journal, Volume 9, Issue 8, pp. 140-145, 2019.

[15] Jaleel Nazir, Muhammad Waseem Iqbal, Tahir Alyas, Muhammad Hamid, Muhammad Saleem, SaadiaMalik and Nadia Tabassum, Load Balancing Framework for Cross-Region Tasks in Cloud Computing, Computers, Materials & Continua, vol. 70, no. 1, 1479-1490, 2022, https://doi.org/10.32604/cmc.2022.019344.

[16] Shubham Gautam, "Load Balancing Algorithms." [Online]. Available: https://www.enjoyalgorithms.com/blog/types-of-load-balancing-algorithms.

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on August 11,2026 at 09:50:03 UTC from IEEE Xplore. Restrictions apply.