

Performance Comparison of Different Load Balancing Algorithms in Cloud Computing

¹Ganesh V. Gujar, ²Satish R. Devane, ³Ratnadeep R. Deshmukh

¹Dr. B.A.M. University, Aurangabad

ganeshvgujar@gmail.com

²MVP's KBTCE, Nasik

principal@kbtcoe.org

³Dr. B.A.M. University, Aurangabad

rrdeshmukh.csit@bamu.ac.in

ABSTRACT-Cloud computing offers economical, scalable, and instantaneous computing resources to enterprises, allowing them to manage substantial traffic volumes and cater to a multitude of users. But the need for effective load balancing techniques has grown significantly as cloud computing becomes more and more popular. To guarantee the best possible performance, availability, and dependability of apps and services, load balancing is a crucial component of cloud computing. This paper offers a comparative study of different cloud computing technologies and load balancing strategies. We present a performance comparison of software-based load balancing; our analysis compares various service broker policies, such as closest distance, optimized, and reconfigurable, with algorithms such as round robin, throttled, and equally spread. Overall, this paper helps readers understand load balancing mechanisms in cloud computing.

Keywords-Cloud Computing, Load Balancing, Load Balancing Algorithms, Round Robin, Throttled, Equally Spread, loudAnalyst.

1. INTRODUCTION

1.1 Cloud computing

Cloud computing is a model for delivering on-demand computing resources over the internet. It involves providing access to shared computing resources, like software, platform, infrastructure, and many. All the services require minimum effort to manage [1]. Rather than owning and maintaining their own computing infrastructure, users can use the cloud service provider's infrastructure to host their applications, data, and services, paying only for the resources they consume. This provides users with the ability to increase or decrease their computing resources as needed, without having to make significant upfront investments in hardware and software [2].

1.2 Cloud computing services

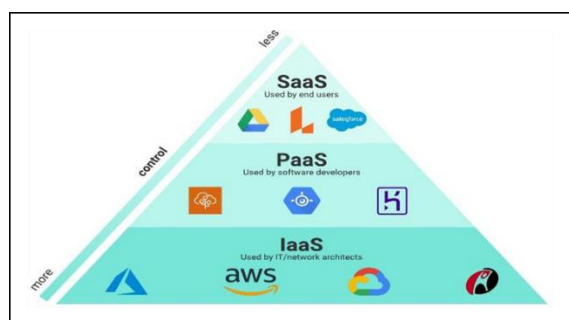


Figure 1. Cloud Computing Services

1.2.1 SaaS

Under the cloud computing concept known as Software as a Service (SaaS), software applications are hosted by a third party and made accessible to users via the internet. Users pay a subscription fee or usage-based pricing to the provider to access the programs and data through a web browser or mobile app, rather than installing and maintaining software on their own PCs or servers. SaaS offers several advantages over traditional software delivery models [3]. Users can access applications from anywhere with an internet connection, without the need for specialized hardware

or software. SaaS providers handle all aspects of software installation, maintenance, and upgrades, reducing the burden on users and freeing up IT resources. Additionally, SaaS offers a scalable and flexible pricing model, with users paying only for the resources they use, without any upfront costs or long-term commitments.

SaaS applications can be used for a wide range of business functions, including customer relationship management (CRM), human resources (HR), accounting and financial management, project management, and collaboration. Popular SaaS providers include Salesforce for CRM, Workday for HR, QuickBooks for accounting, and Microsoft Office 365 for productivity and collaboration.

1.2.2 PaaS

Platform as a Service (PaaS) is a cloud computing model in which a third-party provider offers a platform for customers to develop, run, and manage applications, without the need for infrastructure setup and maintenance. PaaS provides a complete environment for developing, testing, deploying, and managing applications, including hardware, operating system, programming language, and other software components. PaaS users typically pay a subscription fee or usage-based pricing to the provider.

PaaS offers several advantages over traditional application development and hosting models. It allows developers to focus on application development and innovation, rather than infrastructure management. PaaS providers offer scalable and flexible environments that can handle varying workloads, and provide built-in tools and services for application testing, deployment, and management. PaaS also allows for faster time-to-market, as developers can quickly develop and deploy applications without worrying about hardware and infrastructure.

1.2.3 IaaS

Infrastructure as a Service (IaaS) is a cloud computing service where user can access it over the internet. IaaS provides customers with scalable and flexible infrastructure, including virtual machines, storage, and networking, that they can use to run and manage their own applications and workloads. IaaS users typically pay a subscription fee or usage-based pricing to the provider, based on the resources they use.

IaaS provides several advantages over traditional infrastructure deployment models. It allows customers to quickly provide, and scale resources as needed, without the need for physical hardware or infrastructure setup and maintenance. IaaS users have access to a wide range of computing resources, from virtual machines to storage, databases, and networking, and can use them to build and deploy their own applications and services. IaaS also provides users with a high degree of control and flexibility, allowing them to configure and manage their own infrastructure.

IaaS providers offer a range of services, such as virtual machines, object storage, block storage, load balancers, and databases, which customers can use to build and deploy their own applications and services. Some popular IaaS providers include Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), and IBM Cloud.

1.3 Types of Cloud

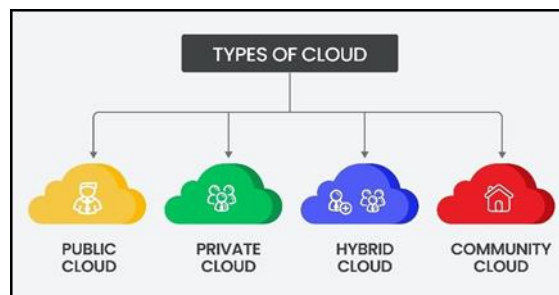


Figure 2. Types of Cloud

1.3.1 Public Cloud: A public cloud is a cloud computing concept whereby computing resources, including virtual machines, storage, and applications, are made available via the internet by a third-party provider [3]. Many customers can access public cloud services, and the cloud provider owns and runs the infrastructure. Google Cloud Platform (GCP), Microsoft Azure, and Amazon Web Services (AWS) are a few examples of public cloud providers.

1.3.2 Private Cloud: A private cloud is a kind of cloud computing where a single company uses dedicated infrastructure. Infrastructure for private cloud services is either managed by the company or by an outside vendor, and services can be hosted on-site or by a third party. Though it involves a larger initial investment and continuous maintenance fees, private cloud offers more security, control, and customization choices.

1.3.3 Hybrid Cloud: A hybrid cloud is a cloud computing model that combines both public and private cloud services, allowing organizations to take advantage of the benefits of both. Hybrid cloud architectures typically involve the integration of on-premises infrastructure with public cloud services, such as cloud storage or computer resources. Hybrid cloud offers greater flexibility, scalability, and cost savings, but also requires additional management and integration complexity.

1.3.4 Community cloud: It is a type of cloud computing deployment model in which a cloud infrastructure is shared by several organizations or entities with a common interest, such as security, compliance, or industry-specific regulations. In a community cloud, the infrastructure is managed and maintained by a third-party cloud service provider, and the participating organizations typically share the costs and resources. This type of cloud deployment model offers benefits such as reduced costs, increased security, and improved scalability [18].

1.3 Load balancing is a key technique for optimizing the performance, availability, and scalability of cloud

applications. Load balancing refers to the process of distributing incoming network traffic across multiple servers, instances, or resources to ensure that no single resource becomes overloaded or overwhelmed.

In the context of cloud computing, load balancing is typically implemented using a load balancer service or software, which sits between the client and the cloud resources. The load balancer monitors the incoming traffic and distributes it across the available resources based on predefined rules and algorithms.

2. LITERATURE SURVEY

2.1 Least Connection: This algorithm directs traffic to the server with the fewest active connections. It is ideal for applications that have long-lived connections, such as web sockets.

2.2 Randomized: This algorithm selects a server at random to handle each incoming request. It is useful for distributing traffic across servers that have similar performance characteristics (5)

2.3 Response Time: This algorithm measures the response time of each server and directs traffic to the server with the fastest response time. It is ideal for applications that require real-time responsiveness.

2.4 RR: Using a cyclic or round-robin approach, the Round Robin load balancing algorithm divides incoming traffic among several servers. A Round Robin algorithm distributes incoming requests among servers in a sequential fashion, giving each one an equal amount.

Maintaining a list of servers and rotating through them in a cyclical manner is how the Round Robin algorithm operates. The method proceeds cyclically, forwarding each incoming request to the subsequent server on the list [6].

A quick and efficient method for distributing the load among several servers is the Round Robin algorithm. It makes certain that every server gets an equal amount of traffic, which helps to keep any server from becoming overloaded. The Round Robin method does not consider the distinct capacities or loads of servers, which may result in an ineffective use of available resources. All things considered, systems with a small number of servers or where all servers have comparable capacities should consider the Round Robin algorithm. Other load balancing methods, including Throttled Load Balancing or Weighted Round Robin, might work better for more complicated systems [7].

2.5 WRR: Weighted Round Robin (WRR) is a kind of load balancing algorithm that divides incoming traffic among

several servers according to their respective weights or capacities. In a WRR algorithm, each server is assigned a weight, which indicates its capacity to handle requests [8]. The WRR algorithm cycles through a list of servers in a round-robin fashion, assigning each server a proportion of traffic according to its weight. The WRR algorithm divides the traffic among the servers according to their weights; for instance, if we have three servers with weights of 8, 4, and 2, our WRR algorithm will assign 8/14 of the traffic to the first server, 4/14 of the traffic to the second server, and 2/14 of the traffic to the third server. After that, the algorithm will cycle through the list of servers once more in a round-robin method.

Because the WRR algorithm considers each server's unique capacity, it offers a more detailed method of distributing the load among several servers. The WRR algorithm lessens the likelihood of overloading servers with lesser capacities by giving each server a weight. This allows more traffic to be directed to servers with higher capabilities.

All things considered, the WRR algorithm works well for dividing up incoming traffic among several servers while taking each one's capacity into account. By maximizing the use of each server, it contributes to the enhancement of distributed systems' dependability and performance. But it doesn't consider varying task durations to assign to the right server.

2.6 TA: Throttled Algorithm [9]: This kind of load balancing algorithm limits the amount of incoming traffic to keep servers from becoming overloaded. This algorithm is frequently applied to distributed systems, which divide up incoming requests across several servers. The fundamental principle of the throttled load balancing algorithm is to keep track of servers, their loads, and their availability—that is, whether they are busy or available. The algorithm routes incoming traffic to the server with the least amount of load after keeping track of each server's current load. But when a certain server's load reaches a certain point, the algorithm reduces the amount of incoming traffic to that server and diverts it to other servers with lesser loads.

In addition to keeping the server from going totally offline, throttling traffic to an overloaded server keeps the overload from impacting other servers in the system. Upon reaching the threshold, the algorithm continues to send traffic to the overloaded server [10].

To balance the load across several servers without overloading any of them, the throttled load balancing algorithm works well overall. Reducing the likelihood of system failures and preventing overloading contributes to enhanced reliability and performance of distributed systems.

Higher level load balancing requirements like Processing Time [11] are not considered.

2.7 Equally spread: Incoming network traffic is dispersed equally among all available resources by this load balancing algorithm in cloud computing [12]. Ensuring that no single resource is overloaded and that all resources are used as efficiently as possible is the aim of this algorithm.

The distribution of incoming traffic in an equally spread load balancing algorithm [13], for each resource, is determined by its current load or utilization. A greater portion of incoming traffic is directed toward resources with lower utilization, and less traffic is directed toward resources with higher utilization. This lessens the chance of overusing or underusing any resource by distributing the workload evenly among them.

Using a dynamic load balancing algorithm, which continuously monitors the utilization of each resource and modifies the traffic distribution, accordingly, is one way to implement evenly spread load balancing in the cloud.

To help balance the workload, the load balancer can, for instance, divert part of the traffic to other underutilized resources if a resource becomes overloaded.

Using a static load balancing algorithm [14], which distributes incoming traffic equally among all available resources at the beginning of the load balancing process [7], is an alternative strategy. Although this method can be easier to understand and more reliable, it might not be as good at managing sudden variations in workload and resource usage.

3 CLOUD ANALYST

An open-source framework called CloudSim is used to model and simulate cloud computing services and infrastructures [15][16]. It enables users and cloud providers to model various cloud environments and assess how well their apps function in various situations.

Cloud services, hosts, data centers, virtual machines (VMs), and other cloud components can all be modeled using CloudSim. To test the functionality of their cloud systems, users can set up the simulation parameters and enter various workloads. Response time, throughput, and resource usage are just a few of the performance metrics that the framework offers for comparing and analyzing the outcomes of various simulation scenarios.

3.1 The primary Features of CloudAnalyst:

1. Support for modeling and simulating complex

cloud infrastructures, including multiple data centers, VMs, and hosts.

2. Support for different cloud services, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

3. The ability to simulate different workloads, including batch processing, parallel processing, and interactive applications.

4. The ability to simulate cloud resource provisioning and allocation algorithms, including load balancing, resource scaling, and scheduling policies.

5. The ability to analyze and compare different performance metrics, including response time, throughput, and resource utilization.

3.2 Simulation Parameters of CloudAnalyst

3.2.1 Region: The geographical area that houses a cloud data center is referred to as a region. A geographical area or a logical collection of neighboring cloud data centers serving a specific user base can be represented as a region. "Regions," which align with the world's six major continents.



Figure 3 Cloud Analyst GUI (Regions)

3.2.2 Users Base: A cloud service provider's user base is the group of users who make requests for computing resources. Individuals, companies, or any other type of entity that uses cloud computing services to carry out tasks or run applications can be considered among these users.

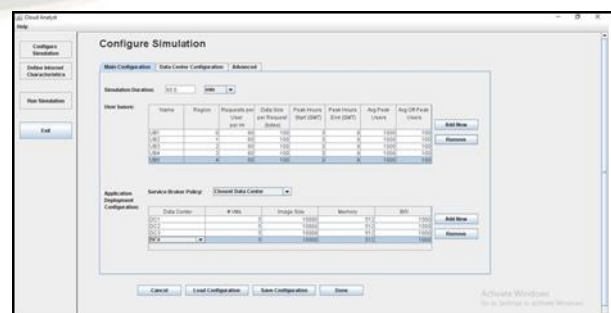


Figure 4: User bases and service broker policy

The user base in CloudSim is modeled and simulated to

evaluate the performance of the cloud computing infrastructure under different scenarios. CloudSim allows users to define the characteristics of the user base, such as the number of users, their arrival patterns, and their workload requirements.

3.2.3 Service Broker: A policy for service brokers is a set of rules or guidelines that control how the service broker acts when choosing cloud services for users. Figure 4 illustrates how the service broker, a software element that functions as a middleman between cloud users and cloud service providers, chooses the optimal cloud service to fulfill the user's needs.

When choosing cloud services, Cloud Analyst's service broker policies aid in optimizing the process by considering various factors like cost, performance, availability, and quality of service requirements. Based on the user's preferences and the existing condition of the cloud infrastructure, the policies give the service broker a framework for decision-making and help them choose the best cloud service.

3.2.4 Data Center controller: A cloud data center's resources and services are managed by a data center controller, a software component that serves as a control entity (Figure 5). Serving as a go-between for the cloud infrastructure and its users, it oversees resource allocation and guarantees the timely and dependable delivery of cloud services.

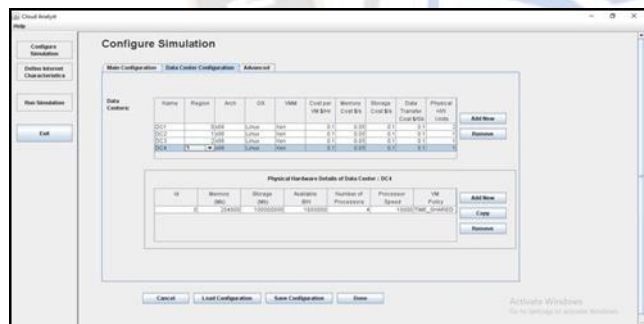


Figure 5: Data Center Configuration

The data center controller in CloudSim provides several functionalities, including:

1. **Resource management:** The data center controller manages the allocation of physical resources, such as CPU, memory, and storage, to the virtual machines (VMs) that are created and destroyed dynamically in response to the user requests.
2. **Scheduling:** The data center controller implements different scheduling policies to manage the workload of the cloud data center. It can prioritize the execution of different tasks or allocate resources based on

different criteria, such as the response time, throughput, or resource utilization.

3. **Service management:** The data center controller manages the delivery of cloud services to the users. It can enforce quality of service (QoS) policies, such as service level agreements (SLAs), and ensure that the services are delivered efficiently and reliably.

4. **Monitoring:** The data center controller monitors the performance of the cloud data center

3.2.5 Internet Characteristics: The class Internet Characteristics represents the features of the internet that serve as a conduit between end users and cloud data centers. Figure 6 illustrates how it simulates cloud service performance over the internet by modeling the network topology, bandwidth, latency, and reliability of the internet.

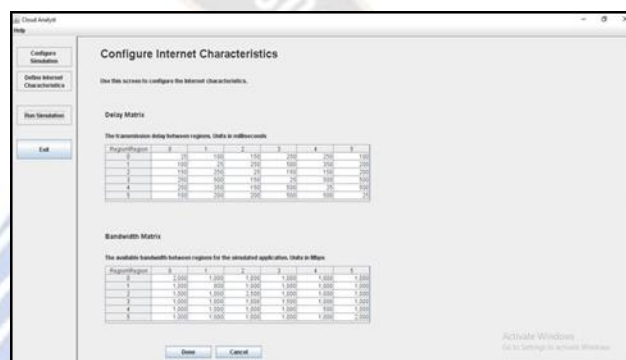


Figure 6: Internet Characteristics

4. PERFORMANCE COMPARISON

Configuring user bases, broker policies, data center configurations, and internet characteristics are among the components of the cloud analyst tool that must be configured to analyze different load balancing algorithms with different service broker policies. As seen in figure 4, we have configured the parameters for each of the six user bases using the region 0, 1, 2, 3, and 4,5 configurations. Figure 5 depicts the data center configuration. It includes four distinct data centers with regions 0, 1, 2, and 3, five virtual machines (VMs) per data center, and it runs the simulation using various service broker policies and load balancing algorithms.

4.1 Performance analysis with RR and closest distance service broker policy

4.1.1 Data Center request serving time: The overall request serving time for each data center for Round Robin algorithm are as shown in table.

Data Center	Avg (ms)	Min (ms)	Max (ms)
DC1	2.318	0.037	4.005
DC2	8.438	0.262	13.264
DC3	13.53	0.197	19.504
DC4	18.683	0.756	25.757

Table 1: Data Center Request Serving Time

4.1.2 User Base Response time: The overall response time and user Base hourly response time by region using RRare as shown in below table.

Userbase	Avg (ms)
UB1	52.167
UB2	58.096
UB3	63.056
UB4	68.121
UB5	315.873
UB6	202.045

4.2 Performance analysis with RR, TA, and ES with Closest distance, Optimized and Reconfigure broker policy: The performance comparison of the load balancing algorithms Round Robin, Throttled, and Equally Spread with the Closest distance, Optimized, and Reconfigure service broker policies is shown in the figures below.

Table 2: User Base Response Time

4.1.3 Simulation Result: The overall results are shown in the below figure.



Figure 7: Simulation Results.

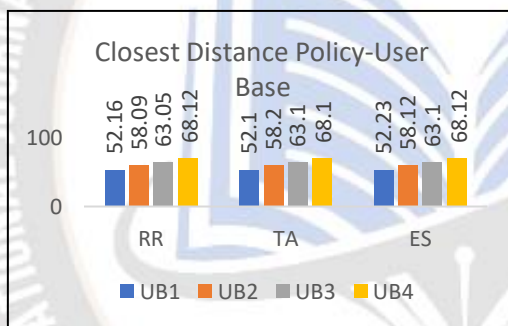


Figure 8: Closest Distance service broker policy-Userbase

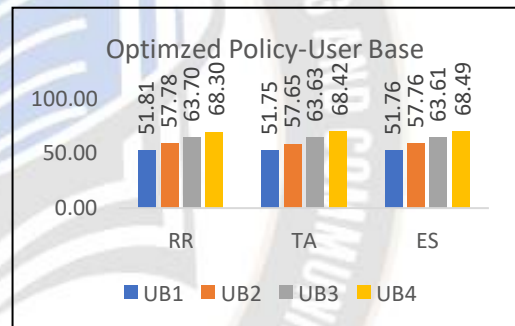


Figure 9: Optimized service broker policy-Userbase

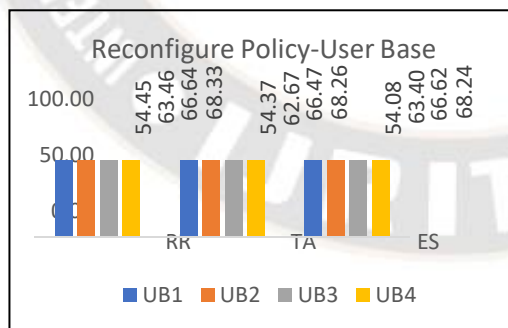


Figure 10: Reconfigure service broker policy-Userbase

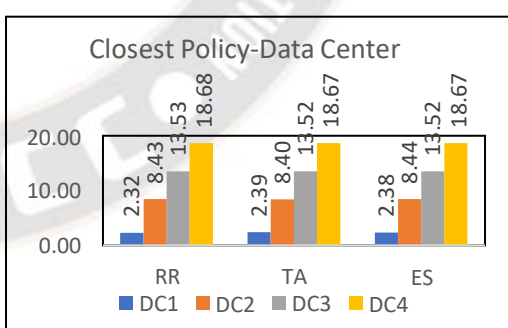


Figure 11: Closest Distance service broker policy-Datacenter.

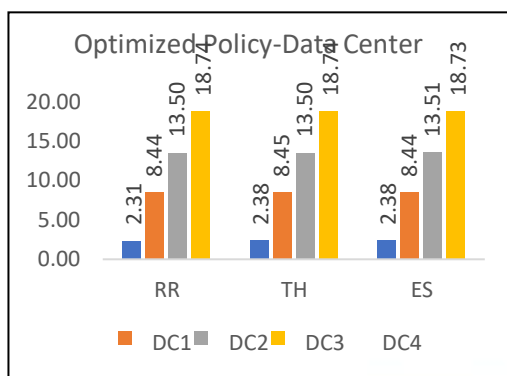


Figure 13: Optimized service broker policy-Datacenter

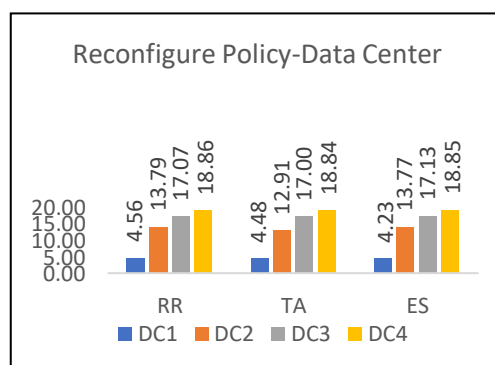


Figure 14: Reconfigure service broker policy-Datacenter.

5. CONCLUSION AND FUTURE WORK

Several load balancing algorithms have been simulated to process user requests in a cloud environment. The scheduling criteria for each algorithm, such as average response time and data center service time of various data centers, are found after analysis. Here, we compared the algorithmic performance using 25VM, 50VM, and 75VM across 4 User Bases and 4 Data Centers. The results are displayed in the graph. When compared to other algorithms, the round robin algorithm performs well overall. Our next task is to create score-based algorithms that will enhance overall execution and response times while being appropriate for heterogeneous cloud environments.

REFERENCES

1. M. Durairaj and P. Kannan, "A Study On Virtualization Techniques And Challenges In Cloud Computing," *Int. J. Sci. Technol. Res.*, vol. 3, no. 11, pp. 147–151, 2014.
2. Sinha, G., & Sinha, D. (2020). Enhanced Weighted Round Robin Algorithm to Balance the Load for Effective Utilization of Resource in Cloud Environment. *EAI Endorsed Transactions on Cloud Systems*, 6(18). <https://doi.org/10.4108/eai.7-9-2020.166284>.
3. Abhinav Chand, N., Hemanth Kumar, A., & Teja Marella, S. (2018). Cloud Computing based on the Load Balancing Algorithm. *International Journal of Engineering & Technology*, 7(4.7), 131. <https://doi.org/10.14419/ijet.v7i4.7.20528>.
4. Kumar, P. (2019). Issues and Challenges of Load Balancing Techniques in Cloud Computing: A Survey. *Issues and Challenges of Load Balancing Techniques in Cloud Computing: A Survey*. February. <https://doi.org/10.1145/3281010>.
5. Priya, V., Kumar, C. S., & Kannan, R. (2019). Resource scheduling algorithm with load balancing for cloud service provisioning. *Applied Soft Computing*, 76, 416-424. doi: 10.1016/j.asoc.2018.12.021.
6. Archana, M., Shastri, M., 2017. A review paper on various load balancing algorithms in cloud computing. *J. Eng. Appl. Sci.* 12 (9), 8579–8585.
7. Kaurav, N.S., Yadav, P., 2019. A genetic algorithm-based load balancing approach for resource optimization for cloud computing environment. *Int. J. Inf. Compute. Sci.* 6 (3), 175–184
8. Rahman, Mazedur & Iqbal, Samira & Gao, Jerry. (2014). Load Balancer as a Service in Cloud Computing. *Proceedings - IEEE 8th International Symposium on Service Oriented System Engineering, SOSE 2014*. 204-211. 10.1109/SOSE.2014.31.
9. Mayur, S., Chaudhary, N., 2019. Enhanced weighted round robin load balancing algorithm in cloud computing, *Int. J. Innov. Technol. Explor. Eng.*, 8(9) Special Issue 2, pp. 148–151, 2019, doi: 10.35940/ijitee.I1030.0789S219
10. Somani, R., Ojha, J., 2014. A Hybrid Approach for VM Load Balancing in Cloud Using CloudSim. *Int. J. Sci. Eng. Technol. Res.* 3 (6), 1734–1739
11. Bhagyalakshmi, Malhotra, M., 2017. A review, different improvised throttled load balancing algorithms in cloud computing environment". *Int. J. Eng. Technol. Manag. Appl. Sci.* 5 (7), 410–416. <https://doi.org/10.26438/ijcse/v6i8.771778>
12. Moharana, S.S., Ramesh, R.D., Powar, D., 2013.

- Analysis of load balancers in cloud computing. *Int. J. Comput. Sci. Eng.* 2 (2), 101–108
16. Falisha, I.N., Purboyo, T.W., Latuconsina, R., Robin, A.R., 2018. Experimental Model for Load Balancing in Cloud Computing Using Equally Spread Current Execution Load Algorithm. *Int. J. Appl. Eng. Res.* 13 (2), 1134–1138.
 17. K. Garala, N. Goswami and P. D. Maheta, "A performance analysis of load Balancing algorithms in Cloud environment," 2015 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2015, pp. 1-6.
 18. Tang, F., Yang, L. T., Tang, C., Li, J., & Guo, M. (2018). A Dynamical and Load-Balanced Flow Scheduling Approach for Big Data Centers in Clouds. *IEEE Transactions on Cloud Computing*, 6(4), 915- 928. doi:10.1109/tcc.2016.2543722
 19. Bhathiya, Wickremasinghe."Cloud Analyst: A Cloud Sim-based Visual Modeller for Analysing Cloud Computing Environments and Applications", 2010, IEEE
 20. R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities," *Proc. of the 7th High Performance Computing and Simulation Conference (HPCS 09)*, IEEE Computer Society, June 2009
 21. P. Gopalakrishnan and B. Uma Maheswari, "Research on enterprise public and private cloud service," *Int.*
 22. *J. Innov. Technol. Explore. Eng.*, vol. 8, no. 6 Special Issue 4, pp. 1453–1459, 2019, doi: 10.35940/ijitee.F1296.0486S419