DOI: https://doi.org/10.52756/ijerr.2023.v30.002 *Int. J. Exp. Res. Rev.*, Vol. 30: 15-24 (2023)

Original Article | Peer Reviewed | Open Access

International Journal of Experimental Research and Review (IJERR)
© Copyright by International Academic Publishing House (IAPH)
ISSN: 2455-4855 (Online)
www.iaph.in

Load balancing techniques in cloud platform: A systematic study

Ranjit Rajak[1*], Anjali Choudhary[1] and Mohammad Sajid[2]

[1]Department of Computer Science and Applications, Dr. Harisingh Gour Central University, Sagar, Madhya Pradesh, India; [2]Department of Computer Science, Aligarh Muslim University, Aligarh-202002, India

E-mail/Orcid Id:

RR, ranjit.jnu@gmail.com, https://orcid.org/0000-0003-2746-3278; *AC*, choudharyanjali202@gmail.com, https://orcid.org/0000-0003-0273-4836; *MS*, sajid.cst@gmail.com, https://orcid.org/0000-0001-8822-5332

Article History:

Received: 24th Jan., 2023
Accepted: 26th Feb., 2023
Published: 30th Apr., 2023

Keywords:
Cloud computing,
load balancing,
makespan,
response time,
virtual machine

Abstract: In the current scenario, researchers have made a new invention and added to the computing paradigm every next second. Cloud computing is one of the most demanding, practical, accessible and extended technologies based on 'pay as per use model' and works on virtualisation via internet. Data sharing and accessing have become easy by properly organising various resources, such as storage, servers, development tools, software, etc, in cloud. Handling these resources has faced many challenges, such as cost management, system performance, migration, load imbalance, reliability and privacy etc. Load imbalance is one of the most important factors which are solved by load balancing techniques. This paper introduced the detailed classification of load balancing methods and techniques that are taken as a solution to overcome such problems and also helps future researchers. Also given is a proposed model for load balancing and some comparative studies of the heuristics methods based on platforms and simulator tools.

Introduction

Currently, the demand for the technological field is growing exponentially in almost all areas and domains. The primary aim is to utilise the processing unit i.e., resources cost-effectively with minimal effort. Cloud computing (Rajak and Rajak, 2021; Rajak, 2018; Sajid and Raza, 2013) provides cost-effective services with minimal human intervention without up-front investment. It is based on the "pay-as-you-go" principles (Sajid and Raza, 2013).

Various organisations like Amazon, Microsoft, IBM, Google, and others provide these services based on end-user payment. These platforms offer private and public services over the Internet (Shafiq et al., 2022). A brief introduction to cloud computing is shown in figure 1 (Shafiq et al., 2022).

The Cloud computing platform provides a backend and a front-end side. The user's primary interaction is the front-end side only, whereas the cloud service model is

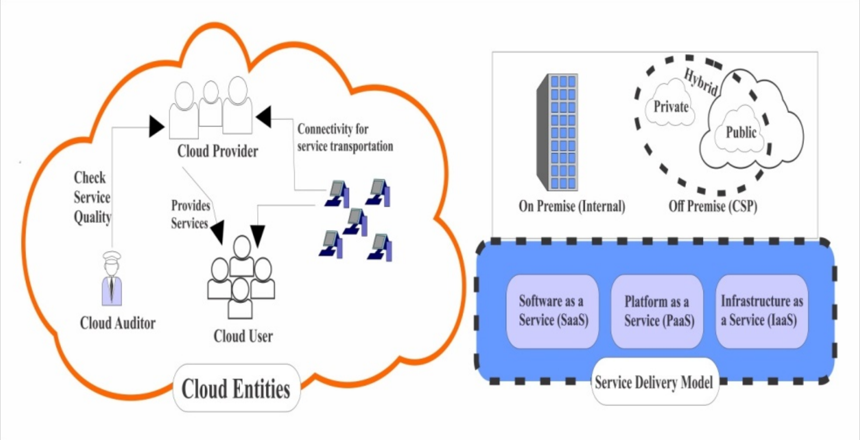

Figure 1 . Overview of cloud computing (Adapted from Shafiq et al., 2022)

***Corresponding Author:** ranjit.jnu@gmail.com

associated with the backend side (Shafiq et al., 2022). The details of these sites and cloud computing architecture are shown in figure 2. Another author (Zhang et al., 2010) stated layer architecture of cloud computing is depicted in figure 3 (Zhang et al., 2010). This architecture is based on four layers: Application, Platform, Infrastructure, and Hardware.

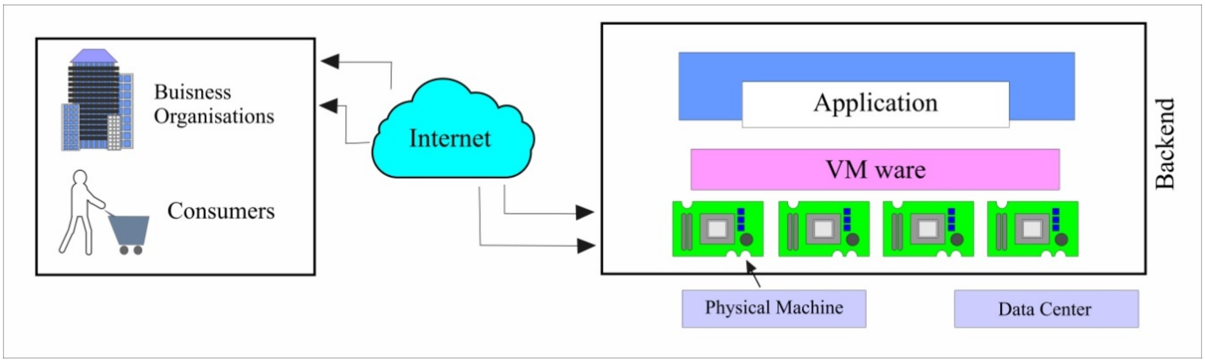

Figure 2. Architecture view of Cloud Computing (Adapted from Shafiq et al., 2022)

Cloud computing is trending for various reasons (Sharma and Sharma, 2021), such as scalability, abstract Infrastructure, support for dynamic and static allocation on a virtual machine, no advanced software and hardware requirement, and no operating system required. Cloud computing supports two models (Rajak, 2018): the deployment model & service model. The private, Public, community and hybrid clouds model belongs to the deployment model. In contrast, software as a service (SaaS), platform as a service (PaaS), and Infrastructure as a service (IaaS) belong to the services model.

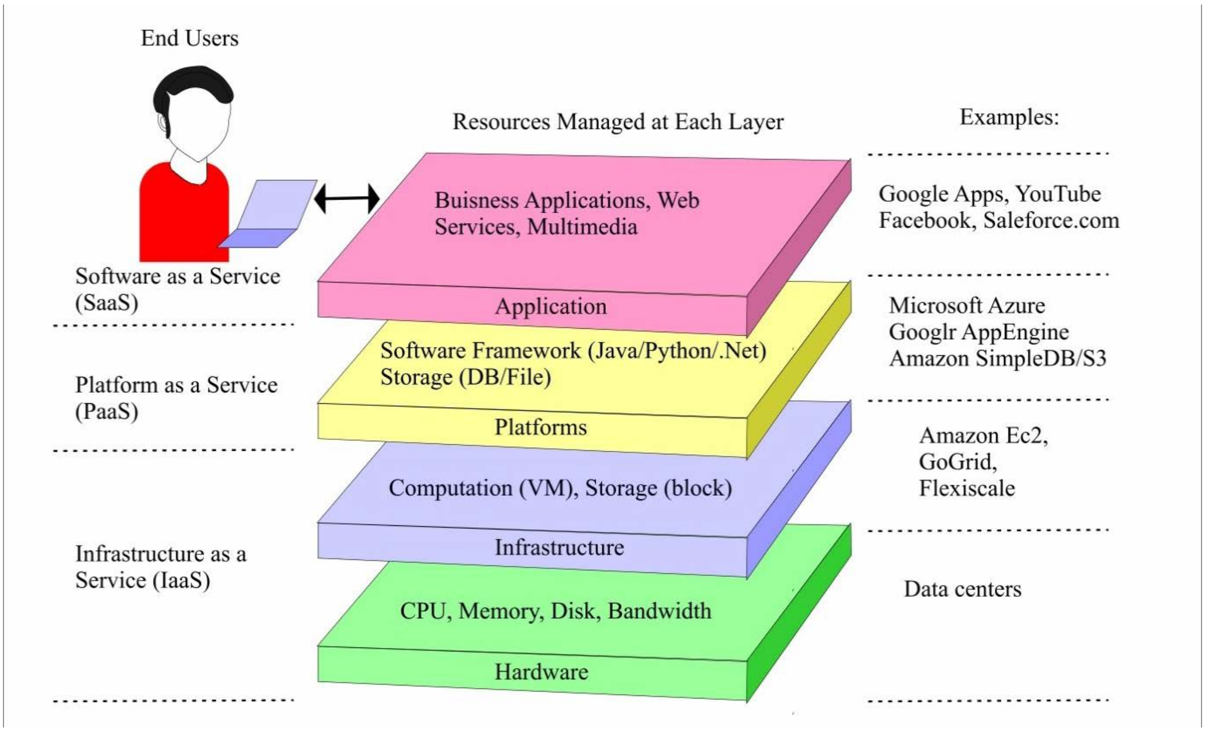
End Users
Resources Managed at Each Layer
Examples:
Buisness Applications, Web Services, Multimedia
Application
Google Apps, YouTube Facebook, Saleforce.com
Software as a Service (SaaS)
Software Framework (Java/Python/.Net) Storage (DB/File)
Platforms
Microsoft Azure Googlr AppEngine Amazon SimpleDB/S3
Platform as a Service (PaaS)
Computation (VM), Storage (block)
Infrastructure
Amazon Ec2, GoGrid, Flexiscale
Infrastructure as a Service (IaaS)
CPU, Memory, Disk, Bandwidth
Hardware
Data centers

Figure 3. Cloud Computing Layers (Adapted from Zhang et al., 2010)

Load balancing (L.B.) is a mechanism to balance the resource utilisation of the virtual machine. It is described as a method to redistribute the workload and resources on the virtual machine for proper utilization. The proper utilization of resources leads to not all the virtual machines being overloaded, the virtual machine being under the load, or the virtual machine being idle. The scheduling (Sajid and Raza, 2016; Shahid et al., 2015) of the tasks on the virtual machine is one factor that affects the balancing of loads of the virtual machines. The load balancing methods' primary job is to identify the next task to be executed to decrease the execution time and optimal utilisation of resources, respectively. The resource allocation and scheduling of the tasks are the two primary load-balancing objectives (Rajak and Rajak, 2021; Sajid and Raza, 2013; Shafiq et al., 2022).

This paper presents the fundamentals of cloud computing and (LB) load-balancing algorithms, followed

DOI: https://doi.org/10.52756/ijerr.2023.v30.002

by the core classification and discussion of load-balancing methods. These algorithms are analysed using brief descriptions, simulators, applications, and environments. This paper also illustrates the various performance parameters such as makespan, execution time, and expected computation time etc.

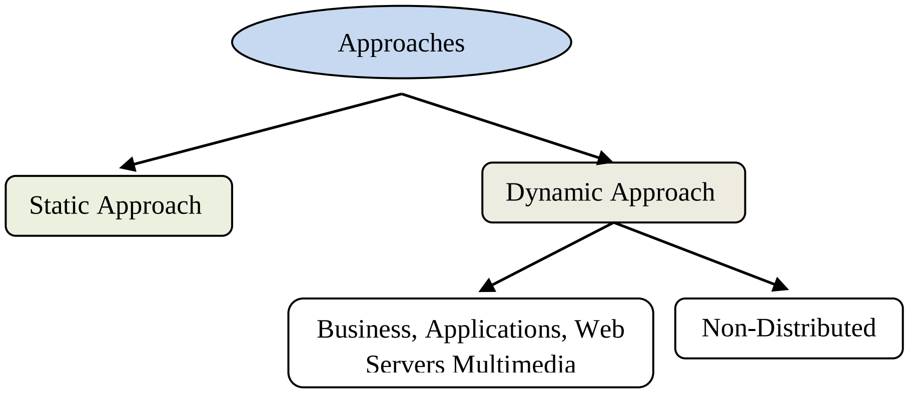

Figure 4. Load balancing Approaches

Section I discusses the fundamentals of cloud computing and load-balancing concepts. Various classification of load-balancing approaches is discussed in Section II. Different types of load balancers are discussed in Section III. Proposed framework of Load balancing is described in Section IV. Performance parameters are discussed in section V. Section VI provides a critical analysis of load-balancing methods based on multiple factors. Conclusion and future scope followed in Section VII.

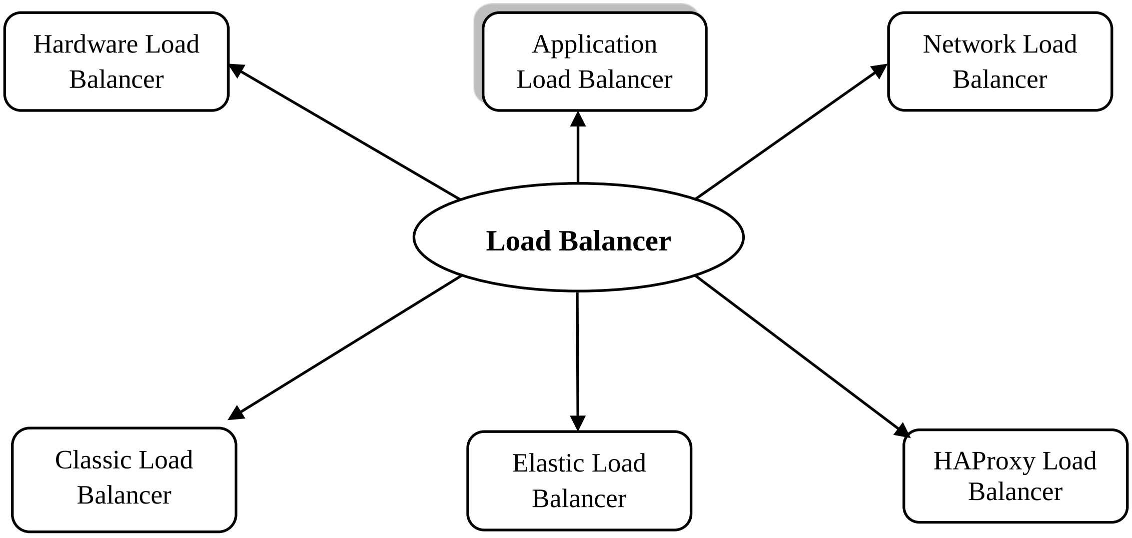

Figure 5. Classification of Load Balancers

Heuristic approaches for load balancing

Load Balancing is the way to assign tasks to the system to utilise the resources better. The fundamental approaches (Zhang et al., 2010; Chaczko et al., 2011) of the load balancing methods are static and dynamic schemes, as shown in figure 4, and brief details are as follows:

Static Approach

This approach requires all information-related load balancing in advance, also called the deterministic approach or compile-time approach. The task traffic is equally distributed among the servers (Darji et al., 2014). This approach is based on two critical factors (Mishra et al., 2020). First is, the idle availability of machines where tasks will be allocated, and the second is the arrival of the tasks at the beginning. The proper utilisation of the resources depends on the schedule of the tasks on the machine.

Dynamic Approach

In this approach, all information related to load balancing is not known in advance, and this information is known at the execution time of the tasks. It is comparatively more flexible than the static approach. The flow of the tasks is not fixed, and the virtual machines will depend on the types of arrival tasks. This approach is divided into two different techniques: distributed and non-distributed. The distributed system allows multiple machines to execute the tasks, and loads are efficiently shared among them. If any machine fails during the execution of the task, it will not stop running the entire process instead of load sharing among the rest of the

DOI: https://doi.org/10.52756/ijerr.2023.v30.002

machines. This condition only causes the system's performance to slow down rather than the halting system. The non-distributed approaches have multiple tasks allocated to only a single machine and execute all tasks. If any kind of fault generates, all systems will be halted.

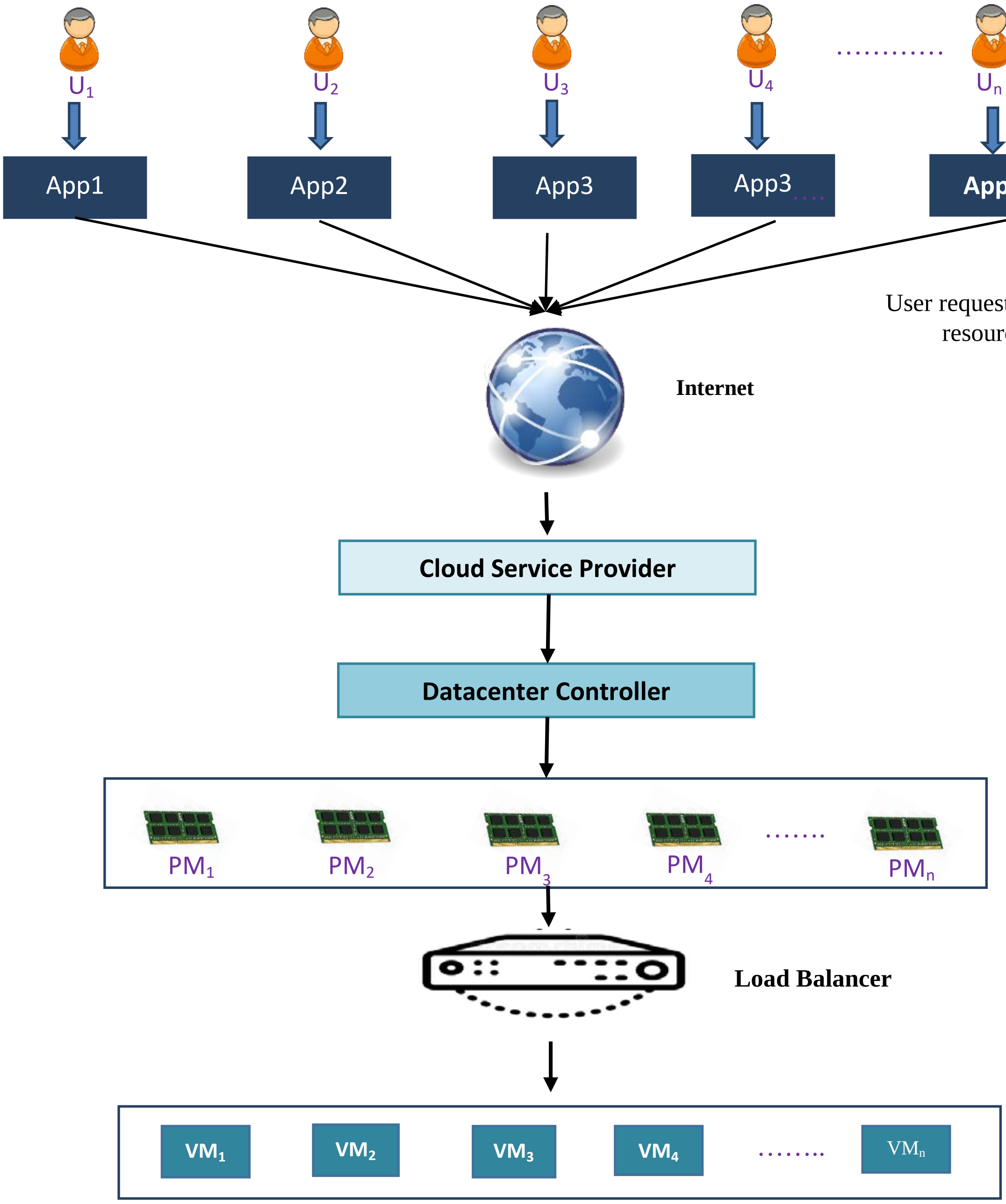

Figure 6. Proposed Framework of Load Balancing

Load balancer

Load balancer is a system which distributes workload across a server and the primary objective of the load balancer is to maintain load onto the system. There are six classifications of load balancers (Mishra et al., 2020) shown in figure 5 as follows:

Hardware Load Balancer (HLB)

This is a physical component of the system and is mainly used to manage each server used in cloud computing. It supports the global load-balancing server and works in a heterogeneous environment and HLB denotes it.

Application Load Balancer (ALB)

It provides service in the seventh layer of OSI model. This load balancer supports both HTTP and HTTPS traffic and ALB denotes it.

DOI: https://doi.org/10.52756/ijerr.2023.v30.002

Network Load Balancer (NLB)

This load balancer works in the fourth layer of OSI and TCP traffic is controlled by this balancer. Primary use of this load balancer in various VM in a cluster as distributor of network traffic.

Classic Load Balancer (CLB)

This balancer works in connection and request levels and is used in Elastic Cloud Compute (EC2).

Elastic Load Balancer (ELB)

It is a combination of three load balancers such as network, classic and application load balancer. i.e., Hybrid load balancer. This load balancer's major job is flooding incoming tasks in various Amazon EC. Hence, an alternative name for this load balancer is AWS.

HAProxy Load Balancer (HAPLB)

HAPLB works in the fourth and seventh layers of the OSI model. Two interfaces are used in HAPLB, which are as follows: one for the user and another for the server LAN. Major uses in reverse proxy and ALOHA Load Balancing.

Proposed framework of load balancing

Here Figure 6 depicts the working model of load balancing. First of all, the number of users who request the resource or virtual machine for scheduling their task by using the application window via the internet. The cloud service provider takes many requests and passes them to the data centre controller. Datacentre controller has all the details of the physical machines and their associated virtual machine. In the next step, load balancers play a vital role and maintain records of which virtual machine is unallocated, or we can say which virtual machine is overloaded, under loaded and idle. Based on this information, it distributed the load among available virtual machines and balanced them efficiently.

Performance parameters for load balancing methods

Throughput

A system can execute user tasks in a given time slot. It is one of the factors for improving the system performance that high throughput leads to the system's high performance (Babbar et al., 2021). Mathematically, it is computed as follows:

$$T^P = \frac{B}{ET} (1)$$

Where B is the size of data in M.B. and E.T. is the elapsed time

Response Time

It is time the system takes when it responds to the end-user after receiving the user request (Babbar et al., 2022). It integrates transmission time, waiting time, and service time, respectively. It is computed as follows:

$$R^{Time} = \frac{T+W+S}{N} (2)$$

Where R^{Time}: Response Time, T: Transmission Time, W: Waiting Time, S: Service Time and N: End-User

Makespan

It is one of the parameters used in a cloud computing platform to evaluate the performance of the methods (Mishra et al., 2020). It considers the maximum completion time of anyjob on the available virtual machine. It is defined as follows:

$$M^{Span} = \mathrm{Max}\left[{E_j}^T\right]_{j=1}^{n} (3)$$

Where M^{Span} : Makespan, n: Number of Virtual Machines

${E_j}^T$: Execution Time on jth Virtual Machine

$${E_j}^T = \sum_{i=1}^{n} \mathrm{X_{ij}} \times \mathrm{ECT_{ij}} \ (4)$$

$$ECT_{ij} = \frac{LT}{PS} \ (5)$$

Where ECT: Expected Time to Compute, L.T.: length of the task, P.S.: Processing Speed.

$$\mathrm{X_{ij}} = \left\{ \begin{array}{c} \text{1 if Task } \mathrm{T_i} \text{ assigned to Virtual Machine V.M.}_{\mathrm{j}} \\ \text{0 if Task } \mathrm{T_i} \text{ not assigned to Virtual Machine V.M.}_{\mathrm{j}} \end{array} \right\} (6)$$

Where X_{ij} is the decision variable

Energy Consumption (E.C.)

E.C. is the most salient parameter in the cloud environment. It consists of the energy amount observed by all I.T. gadgets connected(Mishra et al., 2020). It is also defined as calculating energy consumption in both idle and active states. It is defined as follows:

$$E^{Con.} = \sum_{k=1}^{n} [[{E_j}^T \times \alpha_j + \left(M^{Span} - {E_j}^T\right) \times \beta_j] \times PS \qquad (7)$$

Where VMk consumes αj=joules/MI in an active state,β_j=joules/MI consumed by VM_kin
idle state, PS=Processing speed of kthVM$_k$ in terms of MIPS[a]

Average Communication Time

It stated the average communication time between two consecutive tasks(Kaur et al., 2022)viz., t_a and t_b
can be defined as

$$CT^{Avg_{i,j}} = \frac{\sum_{r_i \in R_i, r_j \in R_j} \mathrm{T}(\mathrm{E}_{\mathrm{t_i}}, _{\mathrm{r_i}} + \mathrm{E}_{\mathrm{t_j}}, _{\mathrm{r_j}})}{|R_i||R_j|} \qquad (8)$$

DOI: https://doi.org/10.52756/ijerr.2023.v30.002

Table 1. Summary of L.B. Algorithm. The table contains eleven previously developed load-balancing algorithms. These algorithms are designated by L^1 to L^{11} and also discussed in a brief explanation of it.

Name of Article	Designated by	Brief Detail
Cloud task scheduling based on load balancing and colony optimization (Li et al., 2011)	L^1	• Based on task scheduling techniques using the LBACO algorithm. • The priority is to balance the whole system's load while optimising the makespan.
Research on the Cloud Computing Load Balance Degree of Priority Scheduling Algorithm based on Convex Optimization Theory (Rong and Bin, 2015)	L^2	• load balance degree of priority scheduling algorithm in cloud computing based on convex optimisation theory using the cluster. • Theoretical and numerical analysis show the effectiveness and feasibility
A novel load-balancing algorithm based on improved particle swarm optimization in a cloud computing environment (Zhu et al., 2016)	L^3	• Improved Particle Swarm Optimisation (PSO) technique used • Performance of each algorithm computed Actual data is generated in exercises and given a demo of the performance.
A proposal for resource allocation management for cloud computing (Alam et al., 2014)	L^4	• Tabu search algorithm is used based on heuristic methods in the cloud environment • Cost constraint, deadline constraint and optimum solution are the factors used in designing a proposed algorithm
An optimal load balancing technique for cloud computing environment using bat algorithm (Sharmaet al., 2016)	L^5	• Introduce load balancing technique based on the Bat algorithm • Used to optimise response time • Balance the load without any latency
Efficient utilization of virtual machines in cloud computing using Synchronized Throttled Load Balancing (Garg et al., 2015)	L^6	• Used Synchronized Throttled Load Balancing algorithm to optimise the load of V.M. • Improve the performance in the cloud platform.
Load balancing algorithm based on estimating finish time of services in cloud computing (Chienet al., 2016)	L^7	• Proposed the technique of calculating the end-of-service time by using a load-balancing algorithm • It takes four scheduling cases to improve the response time and processing time

Where $r_{i,}$ r_j: Subset of resources set R_i and R_j

DOI: https://doi.org/10.52756/ijerr.2023.v30.002

Simulated annealing (SA) based load balancing strategy for cloud computing (Mondal and Choudhury, 2015)	L^8	- used simulated annealing algorithm and compared existing algorithms such as FCFS, R.R. etc - Provide better results as compared to the previous one.
Load balancing in cloud computing: a big picture (Mishra et al., 2020)	L^9	- Discuss various load balancing strategies in different cloud Platform
A genetic load balancing algorithm to improve the QoS metrics for software-defined networking for multimedia applications (Babbar et al., 2022)	L^{10}	- A GLBA is proposed and calculates its parameters. - Focus on the improvement of throughput and reduce response time for users
A deadline-aware load-balancing strategy for cloud computing (Haidriet al., 2022)	L^{11}	- Proposed RDLBS2 strategy, in which migration of incoming request attempts to suitable virtual machines. - Request are referred to as cloudlet and the deadline of these cloudlets are optimise TAT (turnaround time) by utilising unused processing capacity of the VMs

$T(E_{t_i,r_i} + E_{t_j,r_j})$: Execution time of tasks t_i and t_j on resources $r_{i,}$ r_j

Cost

It is the total value of executing and completing the task (Kaur and Kaur, 2022) on a specified virtual machine and is defined as:

$$Cost = \sum(Ca * t) \qquad (9)$$

Ca=Cost per unit time of using resource a concerning t time unit utilised.

Critical analysis of load balancing methods

Load balancing is played a vital in the efficient utilisation of computing resources. Many load-balancing methods have been developed in the cloud computing environment. We have selected some important techniques. These methods are designated by L^1, L^2, L^3, L^4,…L^{11}. We have created three tables (1-3) that briefly discuss the methods, tools used, and their environment that are reviewed and analysed in tabular forms.

Conclusion

Cloud Computing is the most prevalent computing paradigm. It serves the need of different users in education, health, industry, research, government, etc. Cloud computing provides various services such as storage, Infrastructure, software, server, platforms, and networks as per the requirement of end-users and institutions on a pay-as-you-go basis. This paper presents a systematic study of load-balancing algorithms. It discusses limitations, classification, and different challenges of load balancing algorithms. It also critically analyses load-balancing algorithms based on simulators and environments.

The future line is to develop and test the new load-balancing algorithm using different simulation tools and environments. Moreover, the load balancing algorithm can be tested over real cloud platforms like Amazon AWS, Microsoft Azure, etc. Dynamic load-balancing algorithms can also be developed and tested. One possible line of research is to build load-balancing algorithms for an integrated environment consisting of fog computing, IoT, and cloud computing.

Table 2. Comparative study based on simulator used. The table is composed of different simulation tools and programming languages for the implementation of these algorithms. Here 'T' represents yes and 'F' represents no. This table also shows the comparative study of the algorithms based on tools and programming languages used.

Algorithm	Cloud Sim	Cloud Analyst	Grid Sim	MATLAB	Mininet
L^1	T	F	F	F	F
L^2	F	F	T	F	F
L^3	F	F	F	T	F
L^4	T	F	F	F	F
L^5	F	F	F	T	F
L^6	F	T	F	F	F
L^7	T	F	F	F	F
L^8	F	T	F	F	F
L^9	T	F	F	F	F
L^{10}	F	F	F	F	T
L^{11}	T	F	F	F	F

Table 3. Comparative study based on the environment used. The table contains two types of environments that are homogenous and heterogeneous. Here is also compared the eleven algorithms based on the environments.

	Environment	
Algorithm	Homogeneous	Heterogeneous
L^1	T	F
L^2	T	F
L^3	F	T
L^4	F	T
L^5	F	T
L^6	T	F
L^7	F	T
L^8	T	F
L^9	F	T
L^{10}	F	T
L^{11}	F	T

Conflicts of interest

None

References

Alam, M. I., Pandey, M., & Rautaray, S. S. (2014). A proposal of resource allocation management for cloud computing. *International Journal of Cloud Computing and Services Science*, *3*(2), 79.

Amanpreet, K., & Bikrampal, K. (2022). Load balancing optimization based on hybrid Heuristic-Metaheuristic techniques in cloud environment. *Journal of King Saud*

DOI: https://doi.org/10.52756/ijerr.2023.v30.002

University- Computer and Information Sciences, *34*(3), 813-824. https://doi.org/10.1016/j.jksuci.2019.02.010

Babbar, H., Rani, S., Masud, M., Verma, S., Anand, D., & Jhanjhi, N. (2021). Load balancing algorithm for migrating switches in software-defined vehicular networks. *CMC-Comput Mater Continua*, *67*(1), 1301–1316. https://doi.org/10.32604/cmc.2021.014627

Babbar, H., Parthiban, S., Radhakrishnan, G., & Rani, S. (2022). A genetic load balancing algorithm to improve the QoS metrics for software defined networking for multimedia applications. *Multimedia Tools and Applications*, *81*(7), 9111-9129. https://doi.org/10.1007/s11042-021-11467-x

Chaczko, Z., Mahadevan, V., Aslanzadeh, S., &Mcdermid, C. (2011). Availability and load balancing in cloud computing. IACSIT Press.In *International Conference on Computer and Software Modeling, Singapore*, *14*, 134-140.

Chien, N. K., Son, N. H., & Loc, H. D. (2016). Load balancing algorithm based on estimating finish time of services in cloud computing. IEEE. In *2016 18th International Conference on Advanced Communication Technology* (ICACT), pp. 228-233. https://doi.org/10.1109/ICACT.2016.7423340

Darji, V., Shah, J., & Mehta, R. (2014). Survey paper on various load balancing algorithms in cloud computing. *International Journal of Scientific & Engineering Research*, *5*(5), 583-588.

Garg, S., Dwivedi, R. K., & Chauhan, H. (2015). Efficient utilization of virtual machines in cloud computing using Synchronized Throttled Load Balancing.IEEE. In *2015 1st International Conference on Next Generation Computing Technologies* (NGCT), pp. 77-80. https://doi.org/10.1109/NGCT.2015.7375086

Haidri, R. A., Alam, M., Shahid, M., Prakash, S., & Sajid, M. (2022). A deadline aware load balancing strategy for cloud computing. *Concurrency and Computation: Practice and Experience*, *34*(1), e6496. https://doi.org/10.1002/cpe.6496

Li, K., Xu, G., Zhao, G., Dong, Y., & Wang, D. (2011). Cloud task scheduling based on load balancing ant colony optimization. IEEE. In *2011 Sixth Annual China Grid Conference*, pp. 3-9. https://doi.org/10.1109/ChinaGrid.2011.17

Mishra, S. K., Sahoo, B., & Parida, P. P. (2020). Load balancing in cloud computing: a big picture. *Journal of King Saud University-Computer and Information Sciences*, *32*(2), 149-158. https://doi.org/10.1016/j.jksuci.2018.01.003

Mondal, B., & Choudhury, A. (2015). Simulated annealing (SA) based load balancing strategy for cloud computing. *International Journal of Computer Science and Information Technologies*, *6*(4), 3307-3312.

Rajak, R. (2018). A comparative study: Taxonomy of high performance computing (HPC). *International Journal of Electrical and Computer Engineering*, *8*(5), 3386-3391. https://doi.org/10.11591/ijece.v8i5.pp3386-3391

Rajak, N., & Rajak, R. (2021). Performance Metrics for Comparison of Heuristics Task Scheduling Algorithms in Cloud Computing Platform. *Machine Learning Approach for Cloud Data Analytics in IoT*, pp.195-226. https://doi.org/10.1002/9781119785873.ch9

Rong, W., & Bin, R. (2015). Research on the Cloud Computing Load Balance Degree of Priority Scheduling Algorithm based on Convex Optimization Theory. Atlantis Press.In *2015 Conference on Informatization in Education, Management and Business* (IEMB-15), pp. 156-160. https://doi.org/10.2991/iemb-15.2015.31

Sajid, M., & Raza, Z. (2013). Cloud computing: Issues & challenges. In *International*

Conference on Cloud, Big Data and Trust, *20*(13),13-15. https://doi.org/10.1201/b16318-3

Sajid, M., & Raza, Z. (2016). Energy-aware stochastic scheduling model with precedence constraints on DVFS-enabled processors. *Turkish Journal of Electrical Engineering and Computer Sciences*, *24*(5), 4117-4128. https://doi.org/10.3906/elk-1505-112

Shafiq, D. A., Jhanjhi, N. Z., & Abdullah, A. (2022). Load balancing techniques in cloud computing environment: A review. *Journal of King Saud University-Computer and Information Sciences*. *34*(7), 3910-3933. https://doi.org/10.1016/j.jksuci.2021.02.007

Shahid, M., Raza, Z.,& Sajid, M. (2015). Level based batch scheduling strategy with idle slot reduction under DAG constraints for computational grid. *Journal of Systems and Software*, *108*, 110-133. https://doi.org/10.1016/j.jss.2015.06.016

Sharma, S., Luhach, A.K., Abdhullah, S.S. (2016). An optimal load balancing technique for cloud computing environment using bat algorithm. *Ind. J. Sci. Technol.*, *9*(28), 1–4. https://doi.org/10.17485/ijst/2016/v9i28/98384

Sharma, S., & Sajid, M. (2021). Integrated fog and cloud computing issues and challenges. *International Journal of Cloud Applications and Computing* (IJCAC), *11*(4), 174-193. https://doi.org/10.4018/IJCAC.2021100110

Sharma, V., & Sharma, H.C.A. (2021) Review of cloud computing scheduling algorithms. *International Journal of Innovative Science & Research Technology*, *6*(12), 565-570.

Zhang, Q., Cheng, L., &Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. *Journal of Internet Services and Applications*, *1*, 7-18. https://doi.org/10.1007/s13174-010-0007-6

Zhu, Y., Zhao, D., Wang, W., & He, H. (2016). A novel load balancing algorithm based on improved particle swarm optimization in cloud computing environment. Springer International Publishing. In *Human Centered Computing: Second International Conference, HCC 2016, Colombo, Sri Lanka, January 7-9, 2016, Revised Selected Papers 2*. pp. 634-645. https://doi.org/10.1007/978-3-319-31854-7_57

How to cite this Article:

Ranjit Rajak, Anjali Choudhary and Mohammad Sajid (2023). Load balancing techniques in cloud platform: A systematic study. *International Journal of Experimental Research and Review*, *30*, 15-24.

DOI：https://doi.org/10.52756/ijerr.2023.v30.002