A Survey of Load Balancing Methods in Cloud Computing Environments

Snehal Sambhaji Kolte[1] ,Department of Computer Engineering, AISSMSCOE,Pune

Madhavi Ajay Pradhan[2],Department of Computer Engineering, AISSMSCOE,Pune

Abstract: Cloud computing is an emerging model for providing services over the Internet. The load balancing problem is critical in cloud computing for successfully managing cloud resources. Distributing network traffic and computing activity over multiple servers improves resource utilisation, increases throughput, and reduces response time. Load balancing is essential for cloud systems to ensure high availability. Effective task scheduling requires adapting to changing environments and balancing workloads in the cloud. Cloud computing is a high-utility software that has the potential to transform the IT sector and increase its appeal. Proper load balancing is crucial to prevent data loss due to node failures. The study compared different load-balancing strategies.

Keywords: Cloud Computing, Load Balancing, Static Load balancing, dynamic load balancing

1. Introduction:

In the area of computers, cloud computing is a contemporary technology that enables clients to receive services at any time. Resources in a cloud computing system are dispersed globally to provide quicker service to customers. Distributed computing software allows multiple nodes or machines to share computational resources. Recent technical breakthroughs have increased the computing demands of today's software applications, which can outperform current hardware. This category contains applications that need complex computational approaches or involve massive amounts of data, such as Deep Machine Learning, Modelling & Simulation, and other disciplines. Cloud computing offers a practical and efficient alternative for increasing processing capability [1]. New high performance architectures contain specialised hardware for parallel code acceleration, such as Field Programmable Gate Arrays and General-Purpose Graphics Processing Units. Such techniques are efficient in terms of energy and execution time in many scientific applications, but they also induce heterogeneity due to the specialised hardware[2]. The potential rise of cloud computing technology, which addresses large-scale problems, has been facilitated by the advancements in communication technology and increased internet usage. Through cloud computing, users can access hardware and applications as resources over the internet. A route to utility computing, which holds sway over numerous software sectors, is via cloud computing[3]. The cloud service providers will be renting these resources to cloud customers, the cloud resources are kept in a virtual format[4]. In general terms, the cloud's load balancing technique can use either the actual servers or the virtual machines that reside on those servers. The goal of the balancing technique is to evenly distribute the workload across the actual servers or virtual machines. Two approaches can be used for load balancing both dynamic and static load distribution[5]. While dynamic load balancing works effectively in both homogeneous and heterogeneous environments, static load balancing is best suited for homogeneous environments[6][7]. Throughput and use of resources

should rise as a result of the virtual machine's improved load balancing. Load balancing speeds up the completion of operations and concurrently improves system performance[8]. Aside from that, further difficulties include security, increased communication latency, and data loss. Additionally, the burden is reduced and work can be assigned in an optimised manner using the optimisation process. Furthermore, in a cloud environment, the quantity and length of jobs might change quickly, making it challenging to measure every task and carry out the best resource mapping. The main components of load balancing and cloud traffic management are covered in this paper, along with a summary of the load balancing literature and Performace imrovements. Important directions for future research are also covered. These days, cloud services are widely used as a crucial part of smart gadgets and high-end software. Demand is driving up cloud resource utilisation, which is happening daily. Wider domains are linked with cloud computing approaches to store data in several formats. Multicore CPUs enable true multitasking, which enables users to complete more tasks faster by executing numerous complex tasks in concurrently. Multicore processors provide better performance and cutting-edge capabilities that enable systems to operate more efficiently and at lower temperatures. Multicore processors combine two or more processor cores into a single chip. Task Cloud-based scheduling methods are primarily concerned with keeping a steady load while accounting for system bandwidth. This is done to reduce the amount of time needed to finish the task and to increase the productivity and utilisation of the processors[11]. Different The two most popular types of load balancing techniques are fixed and dynamic. In contrast to dynamic load balancing, which rebalances tasks while an algorithm runs, static load balancing pre-assigns tasks to processing parts[23].

1.1 Classification of Load Balancing Techniques

A. Static LB Algorithms: Static load balancing techniques divide up traffic without taking into account the servers' or the system's current condition[9]. Certain static algorithms distribute the same amount of traffic among all servers in a group, either randomly or in a predetermined order. Before any execution begins, static load balancing distributes workloads among resources according to predetermined standards, including task size, complexity, or priority, or resource capacity, availability, or location. Static load balancing has many disadvantages while being easy to use, quick, and reliable. It is not able to adjust to variations in the execution's workload or resource conditions, including demand swings, failures, or heterogeneity. Additionally, it necessitates prior knowledge—which might not be accurate or readily available of the workload and resource characteristics.

The Features of static algorithms are:

1. They require prior understanding about the system.
2. They make decisions based on established rules, such as input load.
3. They are inflexible.

B. Dynamic Load Balancing Algorithms: When making a decision, it considers the information about the present condition of the system[10]. It works better in cloud environments and distributed systems. It evaluates a combination of mastery based on all the information gathered about the nodes and different items of selected nodes, as well as the processes and tasks performed on that node in the cloud. It assigns the

assignment, and under certain circumstances, it must assign it once more using calculations and information gathered. It is challenging to put these algorithms into practice. During execution, dynamic load balancing modifies the workload distribution among resources according to the system's current condition, including workload demand, resource utilisation, or performance feedback. Although dynamic load balancing has several drawbacks, it is more resilient, adaptable, and responsive. Monitoring, analysing, and redistributing the workload necessitates increased coordination, communication, and computational overhead. Additionally, it adds to the complexity and ambiguity of the system's behaviour, which could have an impact on accuracy and performance.

The features of Dynamic algorithms are:

1. They enhance the performance of the system.
2. They make decisions based on the present status of the system.
3. They are adaptable

C. Environmentally conscious Load Balancing: Environmentally conscious load balancing, like the honey-seeking algorithm employed by bee colonies, represents both human and biological nature's activities. How to properly illustrate the best load balancing strategy while working with cloud computing.

2. Measures of Load balancing:

A. Scalability: It is the capacity of an algorithm to carry out consistent load balancing in the system as the number of nodes increases in accordance with the requirements. The chosen algorithm scales up very well.

B. Migration time: The time needed to move a job from a node that is overloaded to one that is underloaded.

C. Makespan: The maximum completion time or the time at which resources are assigned to a user is determined by this measure .It is the time required to complete every task.

D. Response time: It calculates how long it takes the system in total to process a job that has been submitted.

E. Performance: It gauges the effectiveness of the system following the execution of a load-balancing algorithm.

F. Energy consumption: It calculates energy consumption across all nodes. Load balancing prevents overheating and reduces energy waste by distributing the load across all nodes.

3. Approaches in Load Balancing:

A. **Round Robin algorithm**: Each virtual machine is given a set time slice to complete a task in this instance [24]. Once the time slice has elapsed, another virtual machine is allocated another time slice to complete the task. Each virtual machine receives a time slice for task execution in this manner from the round robin algorithm. Certain nodes are overloaded while others are underloaded as a result of its consistent circular movement. It doesn't take into account the server's current.

B. **Weighted Round robin algorithm**: Depending on the capacity of virtual machines, a weight is assigned to each one. The virtual machine would accept requests in accordance with its allocated weight. Accurate execution time forecasting is exceedingly challenging in a real-time cloud computing environment. As a result, this algorithm is not a wise selection [25].
C. **Min-Min algorithm**: This algorithm selects a number of unassigned tasks. Initially, the least completion time for all tasks is calculated. The task with the shortest completion time is picked and executed until all unassigned tasks are completed [26].
D. **Max-Min algorithm:** It calculates the shortest completion time of all tasks. The VM receives the job with the highest execution time. The Max-Min algorithm runs shortest jobs concurrently with longest jobs [11]. The biggest drawback is that jobs with short execution times have starving issues.
E. **Natured inspired algorithms:** Environmentally responsible load balance, such as the bee colony algorithm used by bees to locate honey, represents both biological and human nature activities. How to truly exhibit the best load balancing strategy while working with cloud computing.

Table 1.Overview and Contrast

Aryicle	***Algorithm***	***Evaluation Parameters***	***Tools Used***	***Performance Improvements***
[27]	proposed LB algorithm	Reduces Makespan and provide efficient resource utilization	Cloudsim	Cosidering more SLA Prametres
[28]	Advanced Request Routing (ARR)algorithm	improved Least Outstanding Request	Amazon Web Service	resource utilization needs to be manage carefully
[29]	Load Balancing with Particle Swarm Genetic Optimization Algorithm to Improve Resource Allocation (LBPSGORA)	Cost,Makespan	Cloudsim,MATLAB	Reliability and response time are taken into consideration
[30]	PBMM algorithm	average waiting time and response time	Cloudsim	resource utilizati may be improved
[31]	Predictive Priority-based Modified	makespan, efficiency, and power consumption	Cloudsim	To improve resource provisioning efficiency for end

	Heterogeneous Earliest Finish Time (PMHEFT) algorithm			users, consider proactive techniques. might be addressed.
[32]	Bio Inspired Improved Lion Optimization Meta-Heuristic Approach	response time, throughput and fault tolerance, task migration,	Cloudsim	Increasing the number of nodes and workloads.need to be considered in future
[33]	Proposed priority-based Algorithm	making recruitment tasks more efficient and speedy.	Virtual Machine	apply it on a real life case
[34]	Scheduling Algorithm	Accuracy,Scalability	SinergyCloud	Insufficient native support for real-world workloads,There is insufficient supervision over SLA infractions.
[35]	Weighted Round Robin Dynamic	average completion	Cloudsim	bandwidth fluctuates immensely and Scalability is less
[36]	Resourse Scheduling	Throughput ,Scalability	virtual web servers and physical servers	Balance and response time is not chosen properly
[37]	Ant colony Optimization	execution time and cost Utilization	Virtual Machine	High Makespan
[38]	Genetic Algorithm Dynamic	Improved Performance Better Throughput	CloudSim	Resource requirement decisions need to be taken properly

Table 1 shows overview and contrast of various load balancing approaches with algorithm, evaluation parameters, tools used and their performance aspects.

4. Conclusion

Balancing workloads among cloud nodes is a key difficulty in today's cloud infrastructures. This paper reviews studies on load balancing, a critical feature of cloud computing. Various criteria for load balancing strategies that should be addressed in future load balancing systems. Based on observation have presented a new classification of load balancing techniques Which are general load balancing category, natured inspired algorithm. Some strategies were analysed in terms of metrics, and the results were summarised in tables.Recommend the following for

future works:(1) Research and analyse recent techniques in our specified categor (2) Evaluate and compare each technique using a simulation toolkit and new metrics.

References:

[1] Mrs. Minal Shahakar, Dr. Surenda Mahajan, Dr. Lalit Patil, "Load Balancing in Distributed Cloud Computing: A Reinforcement Learning Algorithms in Heterogeneous Environment", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 11, Issue 2, 2023

[2] Andrea Giordano, Alessio De Rango, Rocco Rongo, Donato D'Ambrosio, and William Spataro, "Dynamic Load Balancing in Parallel Execution of Cellular Automata", IEEE Transactions on Parallel and Distributed Systems, Vol. 32, No. 2, February 2021

[3] M. Saratchandra, A. Shrestha, The role of cloud computing in knowledge management for small and medium enterprises: a systematic literature review, J. Knowl. Manag. 26 (10) (2022) 2668–2698[3]

[4] Y.S. Abdulsalam, M. Hedabou, Security and privacy in cloud computing: technical review, Future Internet 14 (1) (2022) 11.

[5] N.A. Joshi, Technique for balanced load balancing in cloud computing environment, Int. J. Adv. Comput. Sci. Appl. 13 (3) (2022) 110–118

[6] A. Maurya, M. Mohnish, Load balancing in cloud computing, Int. J. Comput. Biol. Bioinform. 8 (1) (2022) 1–7.

[7] V. Sharma, H.C. Sharma, A review of cloud computing scheduling algorithms, Int. J. Innov. Sci. Res. Technol. 6 (12) (2021) 565–570

[8] Kumar, Pawan and Rakesh, Kumar. (2019) "Issues and challenges of load balancing techniques in cloud computing: A survey." ACM Computing Surveys (CSUR) 5:1-35.

[9] M. A. Oxley *et al*., "Makespan and Energy Robust Stochastic Static Resource Allocation of a Bag-of-Tasks to a Heterogeneous Computing System," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 10, pp. 2791-2805, 1 Oct. 2015, doi: 10.1109/TPDS.2014.2362921

[10] U. I. Minhas, R. Woods, D. S. Nikolopoulos and G. Karakonstantis, "Efficient, Dynamic Multi-Task Execution on FPGA-Based Computing Systems," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 3, pp. 710-722, 1 March 2022, doi: 10.1109/TPDS.2021.3101153

[11]G. Rjoub, J. Bentahar, and O. A. Wahab, "BigTrustScheduling: trust-aware big data task scheduling approach in cloud computing environments," Future Generation Computer Systems, vol. 110, pp. 1079–1097, 2020.

[13] R. Yang *et al*., "Performance-Aware Speculative Resource Oversubscription for Large-Scale Clusters," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 7, pp. 1499-1517, 1 July 2020, doi: 10.1109/TPDS.2020.2970013

[14] A. Mukherjee, P. K. Deb and S. Misra, "Timed Loops for Distributed Storage in Wireless Networks," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 3, pp. 698-709, 1 March 2022, doi: 10.1109/TPDS.2021.3100780

[15] Z. Hu, D. Li, D. Zhang, Y. Zhang and B. Peng, "Optimizing Resource Allocation for Data-Parallel Jobs Via GCN-Based Prediction," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 9, pp. 2188-2201, 1 Sept. 2021, doi: 10.1109/TPDS.2021.3055019

[16] D. Cheng, X. Zhou, Y. Wang and C. Jiang, "Adaptive Scheduling Parallel Jobs with Dynamic Batching in Spark Streaming," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 12, pp. 2672-2685, 1 Dec. 2018, doi: 10.1109/TPDS.2018.2846234

[17] J. Jiang, B. An, Y. Jiang, P. Shi, Z. Bu and J. Cao, "Batch Allocation for Tasks with Overlapping Skill Requirements in Crowdsourcing," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 8, pp. 1722-1737, 1 Aug. 2019, doi: 10.1109/TPDS.2019.2894146

[18] W. C. Ao and K. Psounis, "Resource-Constrained Replication Strategies for Hierarchical and Heterogeneous Tasks," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 4, pp. 793-804, 1 April 2020, doi: 10.1109/TPDS.2019.2945294

[19] E. Hwang, S. Kim, T. -k. Yoo, J. -S. Kim, S. Hwang and Y. -r. Choi, "Resource Allocation Policies for Loosely Coupled Applications in Heterogeneous Computing Systems," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 8, pp. 2349-2362, 1 Aug. 2016, doi: 10.1109/TPDS.2015.2461154

[20] N. Kumar, J. Mayank and A. Mondal, "Reliability Aware Energy Optimized Scheduling of Non-Preemptive Periodic Real-Time Tasks on Heterogeneous Multiprocessor System," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 4, pp. 871-885, 1 April 2020, doi: 10.1109/TPDS.2019.2950251

[21] M. Han, J. Park and W. Baek, "Design and Implementation of a Criticality- and Heterogeneity-Aware Runtime System for Task-Parallel Applications," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 5, pp. 1117-1132, 1 May 2021, doi: 10.1109/TPDS.2020.3031911

[22] J. Meng, H. Tan, X. -Y. Li, Z. Han and B. Li, "Online Deadline-Aware Task Dispatching and Scheduling in Edge Computing," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 6, pp. 1270-1286, 1 June 2020, doi: 10.1109/TPDS.2019.2961905

[23] YuAng Chen and Yeh-Ching Chung, "Workload Balancing via Graph Reordering on Multicore Systems", IEEE Transactions on Parallel and Distributed Systems, Vol. 33, No. 5, May 2022.

[24]J. Prassanna and N. Venkataraman, "Treshold-based multiobjective memetic optimized round Robin scheduling for resource efcient load balancing in cloud," Mobile Networks and Applications, vol. 24, no. 4, pp. 1214–1225, 2019

[25] A. Katangur, S. Akkaladevi and S. Vivekanandhan, "Priority Weighted Round Robin Algorithm for Load Balancing in the Cloud," *2022 IEEE 7th International Conference on Smart Cloud (SmartCloud)*, Shanghai, China, 2022, pp. 230-235, doi: 10.1109/SmartCloud55982.2022.00044

[26] Gupta, S., Rani, S., Dixit, A., & Dev, H.: Features exploration of distinct load balancing algorithms in cloud computing environment. International Journal of Advanced Networking and Applications, 11(1), 4177-4183. (2019)

[27] D. A. Shafiq, N. Z. Jhanjhi, A. Abdullah and M. A. Alzain, "A Load Balancing Algorithm for the Data Centres to Optimize Cloud Computing Applications," in *IEEE Access*, vol. 9, pp. 41731-41744, 2021, doi: 10.1109/ACCESS.2021.3065308

[28] D. S. G. . et. al., "An Auto-Scaling Approach to Load Balance Dynamic Workloads for Cloud Systems", *TURCOMAT*, vol. 12, no. 11, pp. 515–531, May 2021.

[29] Mirmohseni, Seyedeh Maedeh & Javadpour, Amir & Tang, Chunming. (2021). LBPSGORA: Create Load Balancing with Particle Swarm Genetic Optimization Algorithm to Improve Resource Allocation and Energy Consumption in Clouds Networks. Mathematical Problems in Engineering. 2021. 1-15. 10.1155/2021/5575129.

[30] Praveenchandar, J., Tamilarasi, A. An Enhanced Load Balancing Approach for Dynamic Resource Allocation in Cloud Environments. Wireless Pers Commun 122, 3757–3776 (2022). https://doi.org/10.1007/s11277-021-09110-x

[31] Sohani, Mayank & Jain, S.. (2021). A Predictive Priority-Based Dynamic Resource Provisioning Scheme With Load Balancing in Heterogeneous Cloud Computing. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3074833.

[32] Kaviarasan R, Balamurugan G, Kalaiyarasan R, Venkata Ravindra Reddy Y,

Effective load balancing approach in cloud computing using Inspired Lion Optimization Algorithm,

e-Prime - Advances in Electrical Engineering, Electronics and Energy,Volume 6,2023,100326,

ISSN 2772-6711,https://doi.org/10.1016/j.prime.2023.100326.

[33] Shahid A. Al Amal Murayki Alruwaili et al., "Proposing a Load Balancing Algorithm For Cloud Computing Applications", *J. Phys.: Conf*, vol. 1979, no. 1, pp. 012034, Aug 2021.

[34] Daniel G. Lago, Rodrigo A.C. da Silva et al., "SinergyCloud: A simulator for evaluation of energy consumption in data centers and hybrid clouds", *J Simulation Modelling Practice and Theory*, vol. 110, no. , pp. 102329, 2021.

[35] SMoly, M.I., Hossain, A., Lecturer, S., Roy, O., 2019. Load balancing approach and algorithm in cloud computing environment. Am. J. Eng. Res. 8 (4), 99–105.

[36] Chen SL, Chen YY, Kuo SH (2017) CLB: a novel load balancing architecture and algorithm for cloud services. Comp Elect Eng 58:154–160

[37] Muteeh, A., Sardaraz, M. and Tahir, M., 2021. MrLBA: multi-resource load balancing algorithm for cloud computing using ant colony optimization. *Cluster Computing*, *24*(4), pp.3135-3145.

[38] Rekha, P. M.; Dakshayini, M. (2019). Efficient task allocation approach using genetic algorithm for cloud environment, *Cluster Computing*,Volume 22,Issue 4, pp 1241–1251, Dec 2019 .

[39] Seth, S., Singh, N., “Dynamic heterogeneous shortest job first (DHSJF): a task scheduling approach for heterogeneous cloud computing systems”, International Journal of Information Technology, Vol. 11, pp. 653–657, 2019

[40] Lin, W., Peng, G., Bian, X. et al., “Scheduling Algorithms for Heterogeneous Cloud Environment: Main Resource Load Balancing Algorithm and Time Balancing Algorithm”, J Grid Computing, vol. 17, pp. 699–726 2019

[41] Longxin Zhang, Liqian Zhou, Ahmad Salah, “Efficient scientific workflow scheduling for deadline-constrained parallel tasks in cloud computing environments”, Information Sciences, vol. 531, pp. 31-46, 2020

[42] Al-Rahayfeh A., Atiewi S., Abuhussein A., Almiani M., “Novel Approach to Task Scheduling and Load Balancing Using the Dominant Sequence Clustering and Mean Shift Clustering Algorithms”, Future Internet, vol. 11, pp. 109, 2019.

[43] V. Kherbache, E. Madelaine and F. Hermenier, "Scheduling Live Migration of Virtual Machines" in IEEE Transactions on Cloud Computing, vol. 8, no. 01, pp. 282- 296, 2020

[44] Shahid, M.A.; Islam, N.; Alam, M.M.; Su’ud, M.M.; Musa, S. A Comprehensive Study of Load Balancing Approaches in the Cloud Computing Environment and a Novel Fault Tolerance Approach. IEEE Access 2020, 8, 130500–130526

[45] X. Gao, R. Liu, and A. Kaushik, “Hierarchical multi-agent optimization for resource allocation in cloud computing,” IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 3, pp. 692–707, 2021.

[46] L. Yin, J. Sun, J. Zhou, Z. Gu, and K. Li, “ECFA: an efficient convergent firefly algorithm for solving task scheduling problems in cloud-edge computing,” IEEE Transactions on Services Computing, vol. 16, no. 5, pp. 3280–3293, 2023