e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

APPLICATION LOAD BALANCER USING MACHINE LEARNING

Shubham Mishra[*1], Rohit Lotlikar[*2], Manish Vishwakarma[*3], Smruti Patil[*4]

[*1,2,3]Information Technology Department, VPPCOE & VA, Mumbai, Maharashtra, India.

[*4]Guide, Information Technology Department , VPPCOE & VA, Mumbai, Maharashtra, India.

ABSTRACT

For every organization to fulfill its demand for its applications, it is imperative to handle the incoming traffic. Hence a load balancer is used to decide which servers are going to be ready to handle the incoming traffic. This helps to take care of good user experience and takes considerably less time to process the requests. A Load balancer manages the distribution of requests across the servers. Load balancing refers to distributing incoming requests across a group of servers. In this paper we propose an answer which might help prevent overusing some servers while some servers are rarely used. This load balancer which is able to route requests to only that server which has the capacity to process that request using real-time data. Our load balancer will use parameters like current CPU Load, physical memory and available storage to see which server is capable of handling the request.

I. INTRODUCTION

High performance computing applications are affected mainly due to Load imbalance. Numerous simulations such as Molecular Dynamics, Adaptive Mesh Refinement and N-Body simulations, tend to exhibit unpredictable behavior. As the applications grows larger the power it consumes also increases which is a main concern in large-scale computing and wasting computing resources due to load imbalance becomes a problem. The method of distributing the network traffic across the servers is called load balancing. It makes sure that there is not much load on a single node, and hence the work is spread evenly, load balancing improves responsiveness of websites and applications for its users. Load balancing is traditionally a hardware appliance. But they are increasingly becoming software oriented. Our take on load balancer is mainly software because of the use of machine learning technologies. Our load balancer will use the parameters to predict the availability of the servers in the future. Compared to other load balancers which exist in the market ours will be different as it will be smart. The most common algorithms of load balancing are not intelligent. These algorithms assign servers either in sequential manner or based on computing power which is not real-time. Our solution will have an intelligent approach. We will make use of machine learning algorithms to dynamically assign servers based on real-time CPU load and available storage. The algorithm we will be using is Q-learning. It is a model–free reinforcement learning algorithm. In reinforcement Learning an agent learns a set of actions to be performed. It performs a set of actions repeatedly and sees the outcome of the actions. The agent gets feedback based on every action. Positive feedback is received for every action in the right path. The agent also gets negative feedback, which makes the agent understand that it is on the wrong path. The agent learns automatically using feedback. Reinforcement learning does not need any data for training thus data acquisition is not needed. Q-learning finds an optimal policy in the sense of maximizing the expected value of the total reward on every successive step, starting from the present state Q-learning can identify an optimal action-selection policy, given unlimited exploration time and a partly-random policy. “Q” is the expected reward for an action taken in a given state.

II. LITERATURE SURVEY

Load balancing carries on the balancing requests using various different pre-configured algorithm. According to the industry standard there are some algorithms such as Round Robin, Weighted Round Robin, Least Connections Algorithm, Least Response Time Algorithm, IP Hash Algorithm.

According to Muhammad Asim Shahid [1] Load balancing is one of the biggest problems with cloud computing, as overloading a device will lead to terrible results that could create technology obsolete. So, it is the need of the hour for creating effective algorithm for efficient use of resources. Load Balancing is also critical for system efficiency, resource time minimization. In the future the creation of fully autonomous new dynamic Load balancing algorithms will allow better use of resources, minimum make span and improved degree of mismatch, effective task migrations, and minimum time span.

According to Ashraf Roshdy [2] a new tool is introduced that enhances the user throughput by 6% on the high loaded cells and 3% on the overall cells based on machine learning algorithm and mobility load balancing concept. Since physical locations of network sites cannot be optimally selected hence this tool offers an optimum solution of adjusting cell borders for traffic balancing which accordingly enhances customer's quality of service.

According to Anna Victoria C.R Oikawa[3] choosing a Load balancing algorithm statically may give a good performance at an initial moment, but because of the application dynamic behavior the load balancer can become inappropriate throughout the execution. Here an approach is presented where the load balancing algorithm decision is automated at runtime using ML named ADAPTIVELB. ADAPTIVELB can switch between LB algorithms according to the need of the given application present at that instant of time, hence achieving the best performance.

According to Kalpana[4]the author has proposed a modification in the improved genetic algorithm such that the time taken for execution is reduced by a significant time . The speed and reliability of this improved genetic algorithm are high hence the chance that any fault occurs is kept to a minimum. The new enhanced and improved genetic algorithm shows high performance when compared to existing genetic algorithm.

According to K. Samunnisa[5] This paper, effective load balancing was used to reduce overhead cost, and maximize capacity utilization, service intervals and resource utilization, etc.

According to Shahbaz Afzal[6] the problem of load unbalancing is discussed along with the factors that lead to it. The author has discussed that there are still quite many issues open in load balancing which can be bridged in the future by applying more efficient and sophisticated algorithm

According to P.P. Geethu Gopinath[7] they have done implemented two different load balancing algorithms namely Min-Min and Max-Min and studied the output thoroughly. The result of the study shows that Min-Min under performs when compared to Max-Min ,but both the algorithms has thier advantages and disadvantages. And finally it is s concluded that the performance of load balancer is mainly dependent upon the cloud environment.

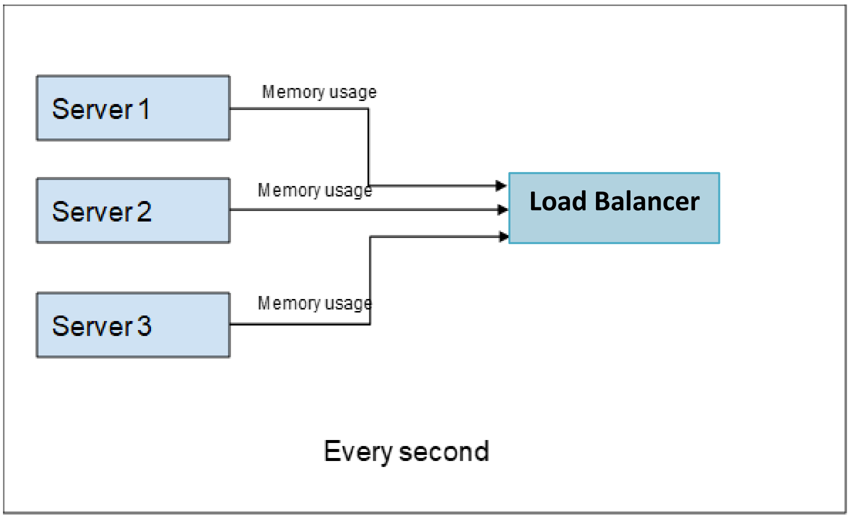
According to Dr.Rajagopalan S [8] to achieve high functionalities and minimum operating cost and time. This article has extensively studdied the different features ,metrics and methods of Server Load Balancer.

III. PROPOSED SYSTEM

In this paper we propose a system based on the Q-learning algorithm. It is a reinforcement learning algorithm. At a given particular state the algorithm learns the value of an action. It rewards the correct action.

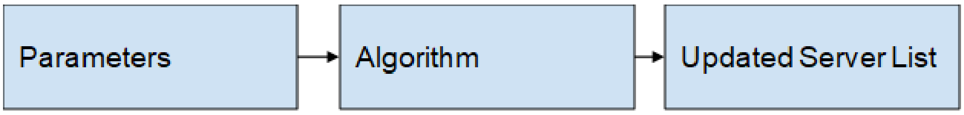
The initial phase involves getting real-time memory usage data from the servers.

- The first step is to get data to feed it to our algorithm
- To achieve this, the servers send real-time memory usage data every second to the load balancer.
- The load balancer has an exposed endpoint to receive this data sent from the servers constantly.



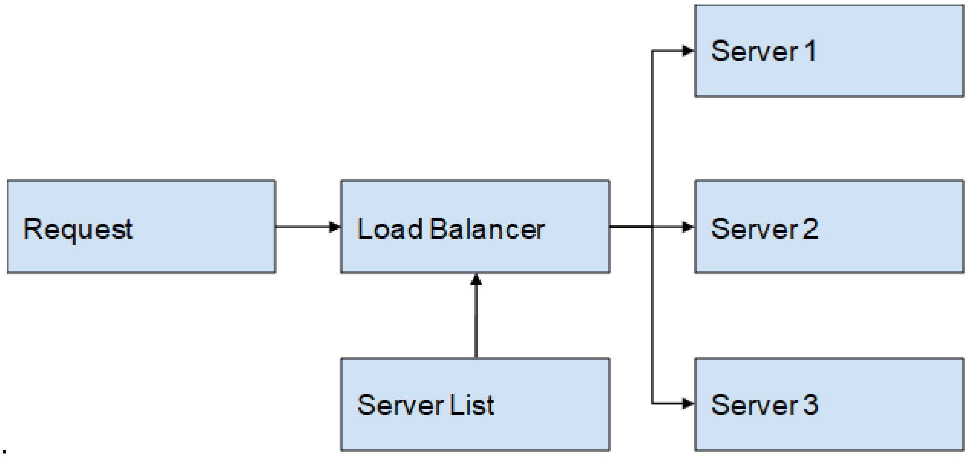
Working of our algorithm.

- The memory usage parameter of the servers is fed to the algorithm.
- Based on the parameters the algorithm updates the list of the servers.



Routing the request.

- Based on the updated server list the load balancer routes the request to the optimal server



IV. CONCLUSION

From our research work, we have reached to a conclusion that smart reinforcement learning algorithm can be used to achieve adaptive load balancing. This method of adaptive load balancing will not only reduce the load on the servers but also make use of the servers efficiently. Out of all the current load balancing method like Round-Robin, Weighted Round Robin etc., this method of load balancing is smarter and more reliable.

V. REFERENCES

[1] M. A. Shahid, N. Islam, M. M. Alam, M. M. Su'ud and S. Musa, "A Comprehensive Study of Load Balancing Approaches in the Cloud Computing Environment and a Novel Fault Tolerance Approach," in IEEE Access, vol. 8, pp. 130500-130526, 2020, doi: 10.1109/ACCESS.2020.3009184.

[2] A. Roshdy, A. Gaber, F. Hantera and M. ElSebai, "Mobility load balancing using machine learning with case study in live network," 2018 International Conference on Innovative Trends in Computer Engineering (ITCE), 2018, pp. 145-150, doi: 10.1109/ITCE.2018.8316614.

[3] C. R. Anna Victoria Oikawa, V. Freitas, M. Castro and L. L. Pilla, "Adaptive Load Balancing based on Machine Learning for Iterative Parallel Applications," 2020 28th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP), 2020, pp. 94-101, doi: 10.1109/PDP50117.2020.00021.

[4] Shanbhog, Manjula & Kalpna,. (2020). Load balancing research paper. 10.35940/ijrte.B1176.0782S619.

[5] Samunnisa, K., Ganesh Kumar and K. Reddy Madhavi. "A Circumscribed Research of Load Balancing Techniques in Cloud Computing." International Journal of Innovative Technology and Exploring Engineering (2019): n. pag.

[6] Afzal, S., Kavitha, G. Load balancing in cloud computing – A hierarchical taxonomical classification. J Cloud Comp **8,** 22 (2019). https://doi.org/10.1186/s13677-019-0146-7

[7] P.P. Geethu Gopinath, Shriram K. Vasudevan,An In-depth Analysis and Study of Load Balancing Techniques in the Cloud Computing Environment, Procedia Computer Science, Volume 50, 2015, Pages 427-432, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.04.009.

[8] S, Dr.Rajagopalan. (2020). An Overview of Server Load Balancing. 2394-9333.