International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 8, August-2018

ENHANCED LOAD BALANCING IN CLOUD COMPUTING

Davender Singh[1] ,Ankur Goyal[2]

[1]*M.Tech Student Computer Science Department YIT college Jaipur*
[2] *HOD Computer Science Department YIT college Jaipur*

Abstract— Load Balancing is a method in which the requests sent to the servers by numerous clients are redirected by the load balancer algorithm/program which seats between the client and the server such that it distributes the load among all the available server instances in such a way that it maximises resource utilization with minimum response time and servers/instances are less overloaded. The rate at which the uses of internet applications at the enterprise level are expanding because of ease of availability of internet and huge set of devices that are available to use internet such as smart phones, laptops, PDAs etc , it challenges the cloud system how to balance the load among the servers present within the cloud environment. These cloud applications services such as SaaS, IaaS and PaaS faces a bottleneck problem when the client increase as a consequence the load on these services increased abruptly so here comes the role of load balancer to balance the traffic by determining the least loaded and fast responding servers from the set of available instances of servers such that it can redirect the request sent from the client. So in order to remain responsive the server instances need to apply some good mechanism of load balancing in the present scenario, enterprises are required to have effective load balancing within their environment architecture. To survive in the competitive market, the wisest solution is to use cloud based service for developing infrastructure. This paper will help in evaluating which load balancing algorithm should be used, using virtual java based cloud environment some dynamic load balancing algorithms are compared with each other and also some enhancement in the already existed load balancing mechanism is proposed and again compared to the existing ones with a thought on overall performance of infrastructure/system and to reduce the response time.

Keywords— load Balancing in cloud, traffic management in cloud computing, dynamic load balancer, comparison of load balancing algorithms, how to choose the best server to reduce the response time, dynamic algorithm of load balancing, weighted load balancing algorithm

I. INTRODUCTION

Users of internet are growing at an alarming rate as now a days the internet services are provided at very cheap prices and all the resources on intent are accessed by various means very easily and handy like smart phones, tablets, personal computers etc.

A study shows that the people who user internet grows rapidly through years as shown by below table

DATE	NUMBER OF USERS	% WORLD POPULATION	INFORMATION SOURCE
December, 1995	16 millions	0.4 %	IDC
July, 2000	359 millions	5.9 %	Nua Ltd.
December, 2000	361 millions	5.8 %	Internet World Stats
March, 2005	888 millions	13.9 %	Internet World Stats
June, 2010	1,966 millions	28.7 %	Internet World Stats
June, 2015	3,270 millions	45.0 %	Internet World Stats
Dec, 2015	3,366 millions	46.4 %	Internet World Stats
Jun. 2016	3,631 millions	49.5 %	Internet World Stats
Dec. 2016	3,696 millions	49.5 %	Internet World Stats
June. 2017	3,885 millions	51.7 %	Internet World Stats
Dec 2017	4,157 millions	54.4 %	Internet World Stats

The reason behind this rapid growth of users of internet are

1. Cheap internet access world wide
2. Can be accessed by various devices
3. Massive internet uses in rural areas

Due to above reasons this leads to flood of request in famous web servers like google , youtube etc. and as the request grows many times the servers become overloaded by these requests and start to respond slow or crashes

Popular examples of sites being crashed due to heavy traffic on it

1. June 25 2009 Michael Jackson crises:- on this date many popular websites faces problem of excess traffic on their servers as it's the date when Michale Jakson passes away. Google on this date faces problem of responding slow while showing the search results, twitter crashed and disables its search results
2. 27 November 2014 bust buy website crashed due to heavy traffic of request on its online shopping site as it's a Thanksgiving weekend. eCommerce sites like amazon, flipkart sales increased about 15% in past year especially in holiday season. eRetail sales for 2 months, just counting purchases made on desktop pc, totaled around $53 Billion, up from nearly $46 Billion a year prior. So the fact is any failure in an online shopping application during peak times is a very serious matter for a retail industry.
3. Oscar Selfie Possibly the most famous site crash that was caused by even more famous people, twelve to be exact. Ellen DeGeneres' selfie taken live at the oscars at the beginning of the month caused so many retweets, that it managed to crash one of the biggest social media sites right now – Twitter. The photo to date has been **retweeted over 3,390,000 times** and counting.

So from the above examples of website being crashed due to heavy traffic and the harm caused by crash we need a solution. In order to server request properly in less time we need a mechanism that will balance the load in servers. This can be achieved by various means

One of the popular solution to balance the load on servers is to replicate the instances of server such that all replica instances of server provides same set of services as the original server is providing but as we replicate the instances of servers and try to balance the load of request on them we face another issue of how we balance the load or distribute the requests on replicas.

The above stated issue is handled by many load balancing algorithms. Load balancing algorithm works between or resides between the client requests and server instances. The main role of load balancing algorithm is to redirect the request from client to the best performing server instance so that the client will get response in no time.

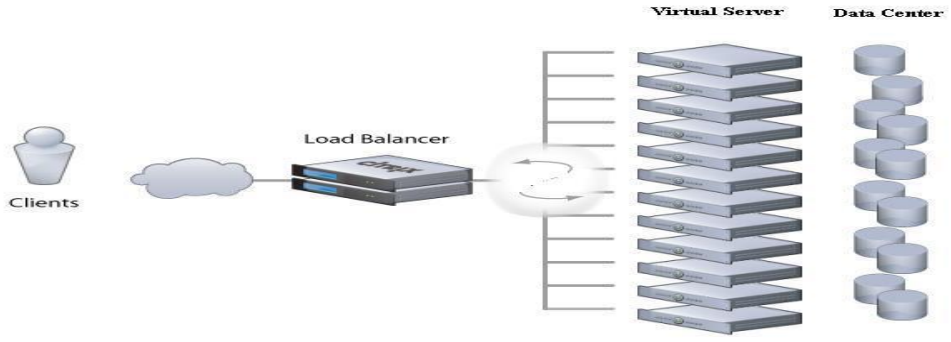


Figure 1: load balancing in cloud computing

[ref]. Load Balancing in Cloud computing, http://community.citrix.com/display/cdn/Load+Balancing

Enhancement in load balancing technique:-

After gone through many research work, found that there is scope of enhancement possible in the current technique of load balancing in order to provide minimum response time or to reduce RTT(Round Trip Time) or latency

The main goal of load balancing mechanism is to select the best server among all the given instances of server so that RTT will reduce.

This paper will propose a different mechanism to balance load in cloud computing environment which will help in reducing latency. In this report the newly proposed load balancer mechanism is compared with the existing and already implemented algorithm of load balancing technique. Comparison is done on the performance of both the algorithm in a virtually created cloud environment using java.

The comparison of proposed algorithm with the already implemented load balancing technique is carried out by placing the load balancing mechanism in between the client and the server and let load balancer chose the best instance for a request from among the available set of instances of server such that the response time minimizes.

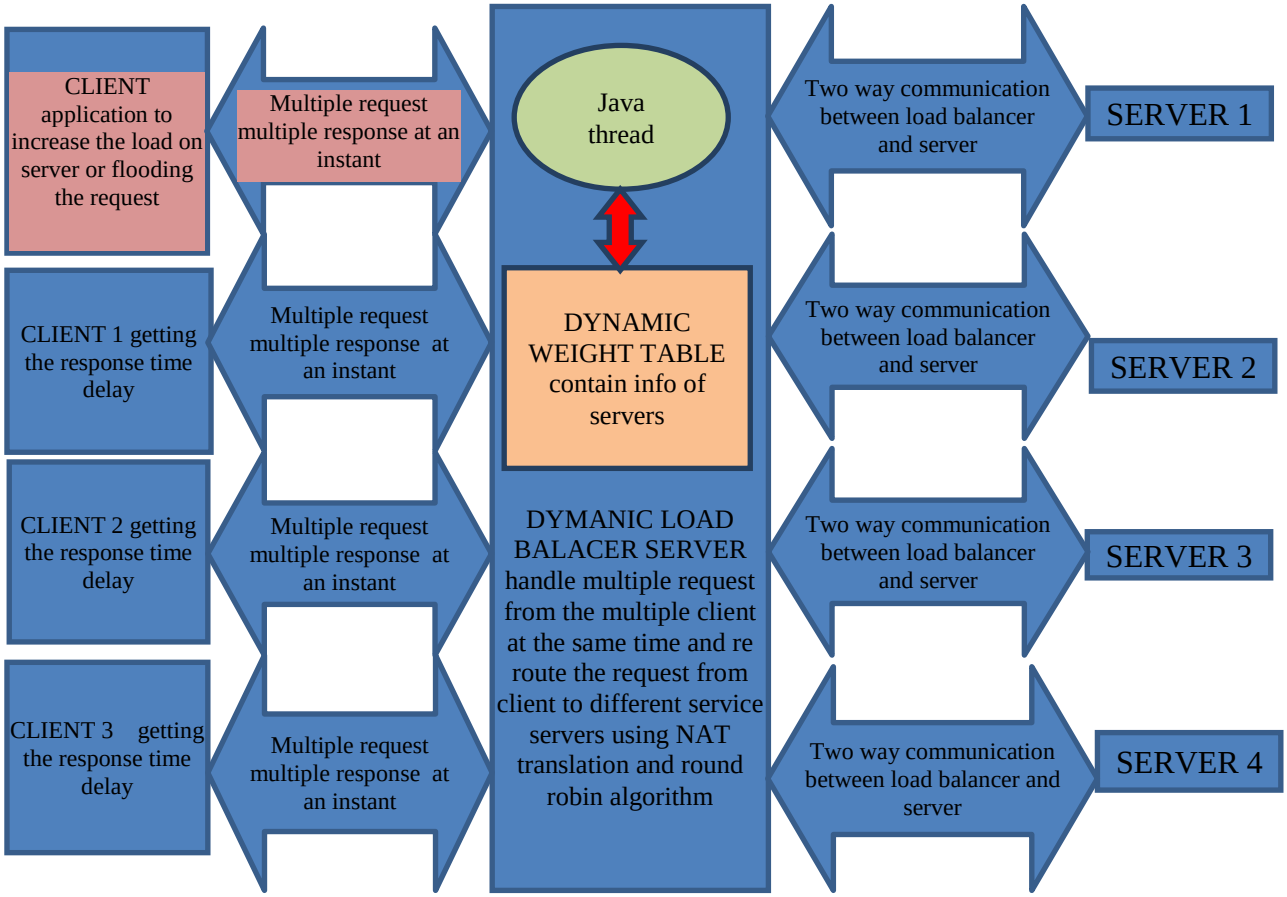
II. METHODOLOGY

Many load balancing algorithm are present to get the job done of selecting the best server by using different mechanisms but in this report we talk about the dynamic load balancing technique which works by combining the logic of least connection and fastest response time.

the algorithm states that The server with less requests and higher throughput time gets the new request. But the disadvantage faced by it is it does not have any weights associated the server, due to which the server starts getting overloaded with requests with time.

As process initiates with the client sending request and then waits for the reposnce the request is then recived by load balancer algorithm and redirected to the best server havening least load and fastest response time then the request is sent to the sected best instance of server by load balancer after that the request is recived and processed by that server and the response is being sent back to the client which is already in the waiting state.

In the above mentioned process the method of selecting the best server by load balancer is based on least connection and having fastest respose time. In our proposed work one extra parameter is included while selecting the best server and thet is the prosessing capability of that server.



Following steps will be followed in the proposed methodology to select the best server instance to reduce the response time and to reduce the latency in our result

Step1. Client will send the request to server asking for a service from cloud

Step 2. Request is being captured by the load balancer algorithm

Step 3. Load balancer algorithm select the best server among the available set of server instances by going through the weighted table rank of servers and select the server having highest rank.

Step 4. The weighted table is maintained by the load balancer algorithm. Which has three fields as load on specific server, its current response time and the processing capability of that server?

Step 5. Rank of all servers is calculated by load balancer with the help of weighted table parameters. Average rank is calculated including all the three parameters by giving the specific weight for each parameter. Best rank server is selected to serve the client

Step 6. After selection of the best server by the load balancer the request is then redirected to that server

Step7. The selected server then processes the client request and responds back.

III. **RESULTS**

So in our proposed work we use the combination of two algorithm which are least connection and least response time algorithm also we introduce the concept of dynamic weight table to chose the best set of servers among all the possible ones at last we select the server which is having the highest rank in weighted table. Ranks are maintained by the load balancer algorithm efficiently

So after applying the algorithm we get the following results with 5 servers in action

RESULTS in tabular form

Number of clients	Original algorithm response time (ms)	Proposed algorithm response time
600	150-200	100 – 174
1500	300-342	200-253

No of clients	Single server with no optimization
1(with one request at a time)	Less than a millisecond
1(with multiple request at a time)	Less than a millisecond
200 clients	20 to 55 ms
600 clients	49 to 123 ms
1500 clients	Above 300 ms
>2000	500 ms to 1 min

IV. CONCLUSIONS

In the proposed solution we have added a extra parameter and modified the original dynamic weighted algorithm. Adding the rank which is correspond to the parameters in weighted table with processing capability of instances to handle the traffic on cloud servers in this we can also extend the work to make the resources to be optimized by considering the hop count, cpu utilization and space management which is also the recent trend in research and development and it is well known in technical world as efficient computing. The future scope of this paper can be enhanced in regards of efficient computing environment

V. REFERENCES

[1] Azzedine Boukerche, Robson Eduardo De Grande, "Dynamic Load Balancing Using Grid Services for HLA---Based Simulations on Large---Scale Distributed Systems," Proceedings of the 2009 13th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications, 2009.

[2] Patrick Wendell, Joe Wenjie Jiang, Micheal J. Freedman, and Jennifer Rexford. DONAR : Decentralized Server Selection for Cloud Services. In Proceedings of the ACM SIGCOMM 2010 conference on SIGCOMM'10. ACM, New Yrk, NY, USA, 231---242

[3] Gowtham Kanagaraj, Naveen Shanmugasundaram And Sathish Prakash "Adaptive Load Balancing Algorithm Using Service Queue" 2nd International Conference on Computer Science and Information Technology (ICCSIT'2012) Singapore April 28---29, 2012

[4] Xiao Qin, Hong Jiang, Yifeng Zhu, David R. Swanson, "Boosting Performance for I/OIntensive Workload by Preemptive Job Migrations in a Cluster System,"

Proceedings of the 15th Symposium on Computer Architecture and High Performance Computing, 2003.

[5] Aameek Singh, Madhukar Korupolu, Dushmanta Mohapatra, "ServerStorage Virtualization: Integration and Load Balancing in Web server," Proceedings of the 2008 ACM/IEEE conference on Supercomputing, 2008.

[6] Hyotaek Lim "Dynamic Load Balancing and Network Monitoring in iATA Protocol for Mobile Appliances", Multimedia and Ubiquitous Engineering (MUE), 2010 4th International Conference on 11---13 Aug. 2010

[7] Cardellini, V. "Dynamic load balancing on Web---server systems" , Internet Computing, IEEE

[8] M.Pathan, C. Vecchiola and R.Buyya, "Load and proximity aware request--- redirection for dynamic load distribution in peering CDNs", in OTM, Nov 2008.

[9] Klaithem Al Nuaimi, Nader Mohamed, Mariam Al Nuaimi and Jameela Al---Jaroodi "A Survey of Load Balancing in Cloud Computing: Challenges and Algorithm" 2012 Second Symposium on Network Cloud Computing and Applications

[10] A. Iosup, S. Ostermann, M. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema. "Performance analysis of cloud computing services for many---tasks scientific computing. In Parallel and Distributed Systems", IEEE Transactions on, 2011. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5719609.