DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5267

Volume 9, No. 1, January-February 2018

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

DETAILED ANALYSIS OF VARIOUS LOAD BALANCING AND POWER REDUCTION METHODOLOGY IN CLOUD ENVIRONMENT

Kamal Narayan Budholia
M. Tech Scholar
Department of computer Science and Engineering
SATI, Vidisha (MP), India

Satish Kumar Pawar
Assistant Professor
Department of Computer Science and Engineering
SATI, Vidisha (MP), India

Abstract: Cloud computing is technology which we are using in our day to day life knowingly or unknowingly. It is like a service which you can take on rental bases as you need. So it reduces operational cost of various industries to a great extent as you don't need to buy all the temporarily needed resources. Load balancing is a major challenge in the field of cloud computing. The objective of the load balancing is to build the trust level of the client by reducing SLA violation. It also provide hike in the resource utilization by fair load distribution. There are various ideas to balance the load in cloud environment. In this paper we analyze various load balancing methodology.

Keywords: cloud computing, Service level agreement (SLA), VM distribution, Virtual machine Consolidation, green cloud.

1. INTRODUCTION

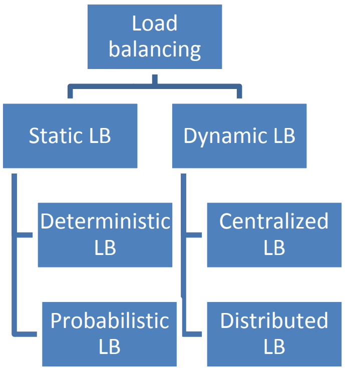
Cloud computing is an idea which provide hardware, Software and platform through the network over the internet on rental bases or pay as you use, in IT companies these services known as Infrastructure as a services (IaaS), Software as a Services (SaaS) and Platform as a Services (PaaS) respectively [1,2]. Many companies like Microsoft, Google, IBM, and Yahoo are the best example of cloud computing service provider through the establishment of data center datacenter at various geographical locations of the world [2]. A datacenter consists of number of host and each host contain a unit called virtual machine. Since practically it is not possible to provide individual resource to each user so we need the concept of sharing resources. To do so we utilize the idea of virtualization. With theconcept of virtualization we can easily implement the main goal of the cloud computing that is to share resources and provide on demand services over the internet in vary efficient way [4].In virtualization we distribute the resources among multiple logical unit called virtual machine (VM). VM act as an interface between hardware and end user and support the facility to run many OS simultaneously [3, 4]. Although there are various issues with cloud computing but in this paper we only address the issue of load balancing. Load Balancing is an idea to distribute the load among other host in the datacenter. Since 75% operational cost of cloud service provider is invested in the form of power datacenter.Green cloud computing is a concept in which we try to minimize energy consumption in turn it minimize carbon emission and operational cost.In this paper we proposed an energy aware dynamic load balancing methodology.

1.1 Service level Agreement (SLA)

It is a contract signed between client and cloud service provider and it also explain about quality of services (QoS) and at any given point services are not up to the point as given in SLA than it is called SLA violation and penalty has to be paid to the customer or end user [7].

1.2 Load balancing

Load balancing is a concept through which we try to evenly distribute load among different nodes in the cloud. Load balancing basically reflects reduction in the SLA violation and increase in effective use of resources by improving efficiency [3]. Load balancing methodology can be categories in following way [6,8]-



Classification of load balancing (LB)

Static load balancing techniques does not consider the running load of the environment instead of the load balancing parameter are fixed throughout the LB operation.

- Deterministic LB –In this load balancing conditions are fixed.
- Probabilistic LB –In this load distribution is depends on certain probability.

Dynamic load balancing techniques are those which distribute the load by considering the running load at a point.

- Centralized LB –The whole responsibility of the load distribution is on a single server or node.
- Distributed LB – Many server contribute to take a decision to balance the load.

Every load balancing methodology works in tow phase, identification phase and selection phase. The identification phase is used to decide whether the host is overloaded or not. After determining the list of the overloaded host we determine on which node we have to transfer the load of the overloaded host, called selection phase[6].

1.3 Server Consolidation

In case, if we are conscious about the operational cost of the data center than we have to concerns about power consumptionas the operational cost of the data center is directly proportional to the power consumption and Energy consumption cost contribute approximately 75% while on 25% cost is needed to operating a data center [11]. To do so we use the concept of server consolidation. In server consolidation we try to minimize the no of server running at a time while keeping in mind the SLA violation. To minimize the no of server running server at a time we shut down all idle server and saves the power consumption, in turn it saves the operational cost of the data center [4]. For that we uses the concept of VM migration.

2 EXISTING LOAD BALANCING METHODOLOGY

In this section we explain and compare the previously discovered load distribution methodology.

The **Minimum Execution Time methodology (MET)** [9] is based on the concept that firstly allocate task to the node that require minimum execution time.Firstly we estimate the execution time of the each task and arrange them inincreasing order of estimated execution time and execute them in order of arrangement. The bottleneck of that methodology is that it does not consider the current load of the nodes which may result to underutilization of some resources and some resource may get overloaded.

The **Minimum Completion Time Methodology** (MCTM) [9] removes the bottleneck of the MET as it also consider the current load of the node by taking node ready time as a load balancing decision parameter. It firstly arrange the all task in increasing order of the completion time of the task. Where the completion time of the task is considered as the addition of execution time of the task and the time the node takes to get ready. Still this methodology suffer from a problem because it calculates the completion time only once as task given to the load balancer.

The **MIN-MIN (MMS) [9]**scheduling methodology works on the problem of MCTM algorithm as the completion time is calculated again and again whenever a task is scheduled. In this algorithm also the task that have minimum completion time at that point get preference over the task that have higher completion time, to execute over the node. In other work the task that require minimum execution time is allocated to the node that provide minimumready time. The problem with this method is starvation. The task that have very high completion time trapped into the starvation situation if the task with smaller completion time arrive again and again.

The **Max-Min**[8] algorithm , in this algorithm the task with higher completion time get preference over the task that have lesser completion time, to get execute. In other word we can say that it execute the task in preference that require maximum execution time and allocate to node that gives minimum ready time. It is a dynamic load balancing algorithm as it updates the parameter timely and consider them in the decision allocation.

The **Equally Spread Current Execution Algorithm (ESCE)** [14], through this concept they explore the idea of equally distribute the load among all. In this algorithm they maintain a table. Each entry of the table contain two values namely the VM id and the number of request given to it. To satisfy request load balancer gives the upcoming request to the least loaded machine. If the situation of same load to two different VM is arrived than resolution is done on FCFS basis. This algorithm treats all node equally so it does not work well in heterogeneous environment.

The **Round Robin (RR) [10]** methodology, in this methodology the load is equally distributed among all the servers, for this, a time span is defined known as quantum time or time slice. Each job done its operation on server for one quantum time than move to another server and work in cyclic manner. This idea improve the response time and decreases SLA violation. Since this technique equally distribute the load among all the server and does not consider the capability of the server, so this this idea does not work well in heterogeneous environment of resources. In case of homogenous resource environment it works well if we only focus on quality of the services but if we consider energy saving also than this algorithm is not good to work on.

The **Weighted Round-Robin (WRR)** [12], this is a dynamic load distribution algorithm and remove the bottleneck of the round robin algorithm by taking server capacity in consideration. It assign weight to each server. Weight basically defined the capacity of the server and according to the weight we assign task to the servers in round robin manner. Higher weighted server get higher load and lightly weighted server get associated with lower load. This algorithm work efficiently in the heterogeneous environment as it also consider the capacity of each server by use of the concept of weight. It also improve the response time but does not consider the concept of server consolidation to reduce the power consumption.

The **Median Absolute Deviation(MAD)** [19], in this methodology the overloaded host are detected to reduce SLA violation. To detect overloaded host they use median absolute deviation formula to calculate the upper threshold by taking 2.5 as a safety parameter and use VM placement policy in which they selects the host to migrates the VM that shows minimum power hike after migration.

The **Inter Quartile Range (IQR)** [19], in this methodology the overloaded host are detected to reduce SLA violation. To detect overloaded host they use inter Quartile Range formula to calculate the upper threshold and lower threshold and use VM placement policy in which they selects the host to migrates the VM that shows minimum power hike after migration. The safety parameter value of inter quartile range method is 1.5.

The **Local Regression (LR)** [19], in this method the future CPU utilization is predicted and use it in overloaded host detection algorithm that in turn reduces SLA violation. To select VM for migration they normally uses minimum migration time policy and in VM placement they place the VM to the host that shows lowest power hike. It uses 1.2 as a safety parameter.

The **ant colony optimization (ACO)** [17], the proposed algorithm in this paper is inspired by the process of food discovery used by ants. In this method of load balancing, in starting they select the node in such a way that it has maximum number of neighboring, called head node. As the ants uses a special kind of liquid called pheromone, in proposed algorithm they also uses the concept of foraging pheromone(FP) and trailing pheromone (FP). Firstly ants starts from the head node and try to find out overloaded node through forward move and keep updating FP trails. In backward move they find under loaded nodes and go back to the over loaded node if this node is still overloaded than they transfer the load of over loaded node to under loaded node and again repeats the process with random neighbors.

The **Honey Bee Foraging Load balancing methodology** [8] [18], it is a distributed load balancing algorithm which is inspired by the bees behavior during food finding process.During food finding there are three type of bees, first scout bees, these bee goes in finding of food and after getting good source of food, returns to the its hive and spread the information about food by means of a special type of activity called, wangle dance, second forager bees goes to the food source informed by scout bees to collect the food and the third one is onlooker bees, those who wait at dancing place.After retuning, forager bees again perform wangle dance to inform about remaining food. The same concept is applied over cloud computing environment where, tasks equivalent to honey bees, Virtual machines behave like food source. Assigning of task to VM is equivalent to the activity of foraging of food source by bees. Over loaded VM is equivalent to the food source that got empty. Than be have to assign task to other VM like discovering other food source. Scout Bees is equivalent to removing task from overloaded VM. As finding suitable VM is like discovering of food source. The activity of informing other bees about the food source by wangle dance is somewhat equivalent to allocating task to under loaded VM and updating the number of task allocated to it and the load of that virtual machine, which in turn will be used by other task to get appropriate VM. This work well with distributed environment.

The **Throttled load balancing (TLB)** [13], in this algorithm they maintain a table which shows the status of the VM (busy/ idle), when a request for the task is arrived at the datacenter, it redirect this request to the load balancer, in turn load balancer go through the table and try to find the VM which can fulfill the need and assign the task and accordingly update the status table, if suitable VM is not available then load balancer queues up the request and when any VM complete its task than status table is updated and these VMs are used to complete waiting queue tasks. It is easy to implement but does not work well for energy saving.

The **Double Threshold Based LB** [15], in this approach they used two parameter called upper and lower threshold is used in order to find overloaded and under loaded nodes. It used best fir strategy with VM migration to balance the load to under loaded nodes and used server consolidation by VM migration to save the power. In this paper they used they defined lower threshold as a static parameter which always fixed to 0.03 while the value of the upper threshold dynamically changes according to the time. Threshold calculation is given as

Lower threshold LT=0.03 and if

U= (MIPS utilization + RAM utilization+ Bandwidth utilization)/3.

Then

Upper Threshold UT=1-x*U.

Where

x=0.05 is an experimental parameter.

The **I Q R basedapproach** [4], in this paper they adopt the idea of double threshold based algorithm, inter quartile range approach of load balancing and proposed an energy efficient approach in which, firstly the calculate upper and lower threshold to categories the hosts in data center. To calculate the threshold they arranger the utilization of the hosts in increasing order and divide the list by median than again calculate median of each list and uses these median as a upper and lower threshold. If at any point the utilization of the host is grater then the upper threshold, they declare them as over loaded host. Best fit strategy based on utilization and power is used to transfer the load of overloaded host to some other host. The lower threshold parameter is used to find out underutilization host. Theyapplied VM consolidation approach to save the energy by transferring the load of underutilized host to some other host by the use of same best fit strategy (take host utilization and power consumption in consideration) and shut down underutilized node after load transfer to save the power.

The **Dynamic VM consolidation approach** [16], they proposed the idea which consists of improved under load decision (IUD) algorithm and minimum average utilization difference algorithm. They defined a threshold which finds out the overloaded node in the datacenter and divides all under loaded nodes in three set. One, which are heavily loaded but not over loaded. Second, which have medium load and third one are those which are very lightly loaded. The load of overloaded host is firstly tried to transfer to the heavily loaded- under loaded node. If node is not available than load is again tried to transfer to medium loaded nodes and again if no medium loaded node is not available than load is finally transferred to lightly loaded machine. In VM consolidation they goes in reverse direction, firstly they try to shut down lightly loaded host then medium loaded host. In this way they perform better in case of energy saving but still have a chance to improve results by modifying threshold calculation method.

3. CONCLUSION

Minimizing operational cost with maintained quality of services is a crucial task in any business. Since operational cost is directly proportional to the power consumption in cloud environment and maintaining quality of services directly affected by load balancing so it is important to load balance the cloud and minimize the power consumption. This paper discusses various method to short out both load balancing and power consumption.

4. REFERENCES

[1] Rajyashree, VineetRichharya "Double threshold based load balancing approach by using VM migration for the cloud computing environment", International Journal of Engineering and Computer Science, Volume 4 Issue 1 January 2015.

[2] SubasishMohapatra, SubhadarshiniMohanty,K.SmrutiRekha, "Analysis of Different Variants in Round Robin Algorithms for Load Balancing in Cloud Computing",International Journal of Computer Applications, Volume 69– No.22, May 2013.

[3] Deepak Mahapatra, Gaurav Kumar Saini, Himanshu Goyal, Amit Bhati, "Ant Colony Optimization: A solution of load balancing in cloud" International Journal of Engineering Applied Science and Technology, Vol. 1 Issue 3, Pages 76-79.

[4] Praveen Shukla, R. K. Pateriya, "I Q R based Approach for Energy Efficient Dynamic VM Consolidation for Green Cloud Data Centers", International journal of Computer Applications, Volume 123- No. 9, August 2015.

[5] Kamyabkhajehei, "Role of Virtualization in Cloud Computing,"International Journal of Advanced Researchin Computer Science and Management Studies (ijarcsms),Volume 2, Issue 4, April 2014.

[6] Sushil Kumar, Deepak Singh Rana, "Various Dynamic Load Balancing Algorithms in Cloud Environment: A Survey", International Journal of Computer Applications (0975 – 8887), Volume 129 – No.6, November2015.

[7] Mohammed Alhamad, Tharam Dillon, and Elizabeth Chang, "A Survey on SLA and Performance Measurement in Cloud Computing", Springer-Verlag Berlin Heidelberg 2011

[8] B S Rajeshwari, Dr. M Dakshayini, "Comprehensive study on load balancing techniques in cloud", an international journal of advanced computer technology, 3 (6), June-2014, Volume-III, Issue-VI.

[9] T. Kokilavani, D.I. George Amalarethinam, "Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing", International Journal of Computer Applications, pp 43-49, Vol 20, No 2, April 2011, DOI: 10.5120/2403-3197.

[10] SubasishMohapatra, SubhadarshiniMohanty, K.SmrutiRekha, "Analysis of Different Variants in Round Robin Algorithms for Load Balancing in Cloud Computing" International Journal of Computer Applications (0975 – 8887)Volume 69– No.22, May 2013.

[11] Anton Beloglazov, and RajkumarBuyya, "Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers", Concurrency and Computation: Practice and Experience, ISSN: 1532-0626, Wiley Press, New York, USA, 2011, DOI: 10.1002/cpe.1867.

[12] Weikun Wang, Giuliano Casale, "Evaluating Weighted Round Robin Load Balancingfor Cloud Web Services" in IEEE 2014.

[13] Durgesh Patel, Mr. Anand S Rajawa, "Efficient Throttled Load Balancing Algorithm in Cloud Environment" International Journal of Modern Trends in Engineeringand Research, 2015.

[14] VishwasBagwaiya, Sandeep K Raghuwanshi, "Hybrid approach using throttled and ESCE load balancing algorithms in cloud computing", in IEEE, 2014.

[15] Rajyashree, VineetRichhariya, "Double Threshold based Load Balancing Approach by using VM migration for the cloud computing Environment" International Journal of Engineering and Computer Science, Volume 4 Issue 1 January 2015, pp. 9966-9970.

[16] Dongyan Deng, Kejing He, Yanhua Chen, "Dynamic Virtual machine consolidation for improving energy efficiency in cloud data center", IEEE, 2016.

[17] Kumar Nishant, KunwarPratapSingh,Pratik Sharma, Vishal Krishna,Chhavi Gupta, Nitin, Ravi Rastogi, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization", in proceedings of 14th International Conference on Computer Modeling and Simulation, pp 3-8, 2012.

[18] Sheeja Y S, Jayalekshmi S, "Cost Effective Load Balancing Based on honey beeBehavior in Cloud Environment", First International Conference on Computational Systems and Communications (ICCSC) in Trivandrum, IEEE, 2014.

[19] Seema Vahora, Ritesh Patel, "CloudSim-A Survey on VM ManagementTechniques", International journal of advanced research in Computer and communication engineering, Vol 4, issue 1, January 2015.