A STUDY ON LOAD BALANCING TECHNIQUES IN CLOUD COMPUTING ENVIRONMENT

Suriya Begum [1], **Kavya Sulegaon**[2] **,Venugopal** [3]

[1] Professor,Department of Computer Science and Engineering,New Horizon College of Engineering, Bangalore, India.

[2],[3] PG Scholar, Department of Computer Science and Engineering,New Horizon College of Engineering, Bangalore, India.

[1]suriyabegumnotes@gmail.com; [2]kavyasulegaon92@gmail.com

ABSTRACT*: *Almost all are facing problems like finding space to store all the information in personal computer. Using cloud computing one can store the information like files, applications videos, music and so on the cloud (Internet), instead of storing them on the personal computer. Almost all the information is stored on servers which are maintained and controlled by cloud providers such as gmail, apple, Microsoft, Google, Amazon and so on. And the number of clients using cloud is also increasing. Cloud computing is used by all the business organisations and for personal use. For a cloud datacenter, the biggest issue is how to tackle billions of requests coming from cloud end users to handle such request efficiently.We need to distribute the load equally among multiple devices/servers in order to avoid burden on single system. To improve the load balancing, various techniques have been proposed by researchers. This paper describes a survey on load balancing schemes in cloud environment. There were various load balancing techniques discussed in this paper and their corresponding advantages, disadvantages and performance metrics are studied in detail.

KEYWORDS*: *Cloud Computing, Cloud Providers, Internet, Load Balancing, Performance Metrics .

I. INTRODUCTION

Load balancing is the process of reassigning the total loads to the individual nodes of the collective system to make the best response time , good utilization of the resources. Simultaneously removing a condition in which some of the nodes are over loaded while some others are under loaded [7] .Cloud Computing is an internet computing in which the load balancing is the one of the challenging task. To make a system better by allocating the loads to the nodes in a balanced manner,various methods are to be used, but due to network congestion, bandwidth usage etc, there were problems which occurred. These problems were solved by some of the existing techniques. To maintain the performance of cloud computing it is necessary to decrease the workload. One solution to overcome such issues is to balance the load by applying load balancing algorithms. Load balancing divides the work of one server into the available servers. In this way more work gets done in the same time[2]. There are two types of load balancing algorithms, static and dynamic. In the Static algorithm, prior knowledge of the system is needed like resources of the system , task details. The execution of these algorithms do not take into account the current state of the system , hence these algorithms do not depend on the current state . Some of the Static algorithms' are Round Robin Algorithm, Randomized Algorithm, Central Manager Algorithm. Dynamic algorithm is based on the current system and it gives better performance than the Static algorithm.[7]. During the execution of the algorithm, workload is distributed among the processors unlike the Static algorithm ,Dynamic algorithm buffers the process in the queue on the main node and allocates dynamically upon request from remote nodes. As a result, Dynamic load balancing algorithms can provide a significant improvement in performance over Static algorithm [3], these algorithms are complex but more fault tolerant and have an overall better performance than Static algorithms, one of the Dynamic algorithms is Parallel Graph Partitioning [1].

Virtualization is an important concept in cloud computing. Virtualization is a process of creating multiple virtual machine instances for a single host machine in order to serve more number of users .Various applications and operating systems runs on same server node at a time. Virtualization technology has incremented the flexibility and security of cloud [4].

There are many issues related to Cloud computing and are increasing day by day. There are many crucial problems associated with it. Load balancing is one such issue. Load balancing means dividing the incoming load or tasks equally among the cloud nodes so as to achieve maximum user satisfaction, reduce response time and maximize resource

doi : 10.17950/ijer/v5i4/023

utilization.Load balancing issue in Cloud Computing can be resolved primarily in two ways :
1. Assigning the tasks to suitable nodes in such a manner that load gets distributed evenly and
2. Migration/Reallocation of VMs .

For the first sub-problem, algorithms has been developed to be applied at load balancer which will decide to which VM a particular request will be assigned. In the second sub-problem, when a particular host gets overloaded, then some of its VMs get migrated to the hosts which are lightly loaded. The policies have been applied to decide which VMs to migrate from a particular overloaded host and to which host among the lightly loaded ones, the VMs should be re-allocated [4]. In our study, various load balancing techniques has been discussed .

II. EXISTING TECHNIQUES IN LOAD BALANCING

Anjali [2] proposed a Dynamic Load Balancing in Cloud Computing using agents. The use of mobile has shown better results than existing load balancing algorithms. The whole working is done with the help of agents known as a regular agent and a mobile agent. Mobile agent is a program which migrates from one machine to another. With the help of this, an executable code is moved to a new host. It runs independently according to the interest of client. Mobile agent adds to regular agent. The different features of mobile agent which makes it unique are, the capability of learning and mobility.
In addition to this, mobile agents also have the benefits of bandwidth conservation, reduction of completion time, load balancing, dynamic deployment, etc. Mobile agent is used for monitoring, information retrieval, remote control and dynamic systems. The proposed approach greatly reduces the communication cost of servers, accelerates the rate of load balancing which indirectly improves the throughput and response time of the cloud.

Reena Panwar [3] has introduced a dynamic load management algorithm for distribution of the entire incoming request among the virtual machines(VMs) effectively. The load is managed by the server by considering the present status of present VMs for request assignment sharply for all free VMs to be used at request assignment and will take more requests that are dynamic in nature Therefore, reduction in response time when compared to VM Assign Algorithm. The experimental results has shown that this algorithm have minimum response time and proper resource utilization by using Cloud Analyst tool and checked its performance on various different load distributions.

Surbhi Kapoor [4] considered a Cluster based load balancing which works well in heterogeneous nodes environment, considers resource specific demands of the tasks and reduces scanning overhead by dividing the machines into clusters. When user submit the task , the load balancer matches task resources specific requirements with capacity range of cluster in order to assign the task to appropriate cluster. Then among the clusters, load balancer matches the suitable available VM to which task must be assigned, scanning of lists is divided into two levels .This will reduce the overhead involved in scanning list and can assign better VM to the task. Advantage of this algorithm less time consumption. Suitable for heterogeneous environment and it also considers resource specific demands of the tasks.

Tejinder Sharma [5] proposed an efficient and enhanced scheduling algorithm that can maintain the load balancing and provides better improved strategies through efficient job scheduling and modified resource allocation techniques. Load Balancing is done with the help of load balancers where each incoming request is redirected and is transparent to client who makes the request. Based on parameters, such as, availability or current load, the load balancer uses various scheduling algorithms to determine which server should handle and forwards the request on to the selected server. Load balancing ensures that all the processors in the system as well as in the network does approximately the equal amount of work at any instant of time. The overall response time and data centre processing time is improved as well as cost is reduced.

Priyank Singhal [6], the Authors have focused on a two level task distribution system over a three tier cloud architecture. They studied the use of a hybrid task scheduling algorithm which combines two commonly used scheduling methods, the MM (Min-Min) and OLB (Opportunistic Load Balancing) to create a hybrid Balanced Load Min-Min algorithm (BLMM) algorithm. The concept of BLMM scheduling algorithm is to distribute task among each service manager into some subtasks to be executed in a suitable service node. BLMM considers the execution time of each subtask on each service node. Each subtask will be figured out the execution time on different service nodes through agent. According to the information gathering by agent, each service manager chooses the service node of shortest execution time to execute different subtasks and records it into the Min-time array. Finally, the Min-time array of each subtask is recorded that is a set of minimal execution time on certain service nodes. This leads to more efficient execution and maintains load balancing of the system nodes.

Suriya Begum et.al.[8] proposed a Mathematical model exclusively considering virtual machine for performing load balancing. The system jointly addresses the routing as well as task scheduling and also focuses on the issues pertaining to resource allocation. A novel mathematical model considering Stochastic model for load balancing and scheduling in cloud computing clusters has been developed . A cloud system consists of a number of networked servers. Each of the servers may host multiple Virtual Machines. Each Virtual Machine requires a set of resources, including CPU, memory, and storage space, considered a stochastic model for load balancing and scheduling in cloud computing clusters. A primary contribution is the development of frame-based non-pre-emptive VM configuration policies. These policies are made nearly throughput-optimal by choosing sufficiently long frame durations, whereas the widely used best fit policy was shown to be not throughput optimal. Simulations indicate that long frame durations are not only good from a throughput perspective but also seems to provide good delay performance.

Neha Gupta [9] has discussed about Genetic Algorithm in Mobile Cloud Computing (MCC). MCC combines the mobile devices and cloud computing is a new platform to create a new infrastructure, where cloud performs the computing tasks and storing enormous amounts of data. In MCC, data processing and data storage happen in the external of the mobile devices. MCC provides various advantages, such as , improvement in data storage capacity and processing power and it improves synchronization of data because they are gathered and stored in one place. MCC is the combination of cloud computing and mobile networks and it is used to bring benefits for mobile users, network operators, as well as cloud computing providers that varies cloud resources and network technologies towards unrestricted functionality, storage, and mobility to provide a gathering of applications on mobile devices anywhere, on the pay-as-you-use service anytime through the channel of internet .The main goal of MCC is to enable execution of mobile applications on a plethora of mobile devices.

In the proposed framework , the authors used genetic based techniques to balance the load of the system and to schedule the request to different virtual machines , by finding the right mapping solution and with reduced response time of execution of requests. A genetic algorithm is a type of searching algorithm. It searches a solution space for an optimal solution to a problem. The key characteristic of the genetic algorithm is the way searching is done.

The Genetic algorithm is composed of four functions:

A) *Population Size:* In this first of all two array lists are created one is for containing the list of virtual machines over which tasks are to be scheduled to execute and other is for containing the list of requests to be processed.

B) *Objective Function:* This function will fetch all the possible virtual machine over which task is to be executed.

C) *Mutation Function:* This function will schedule the task to the resource queue according to the calculation required by the tasks and the ability of resources to process the task or we can say it can find the best virtual machine for task to be executed.

D) *Fitness Function:* This function will call the virtual machine that satisfies the condition by performing the above function when the task is to done.

B.Bhaskar [10] has developed a novel Round Robin Algorithm for load balancing in cloud computing. Here, skewness measurement technique introduced along with load balancing using Round Robin Algorithm , in order to provide enhanced performance which results in Green Computing. Skewness concept is used to measure the utilization rate of a node.Cloud partitioning is the process of dividing a huge public cloud into sub partitions. Each cloud partition contains some number of nodes, one node might be working for a long time while other nodes are sitting idle. Despite a node being utilized for a long time the cloud partition status will be showing normal. This is because of the remaining nodes in that particular partition are sitting idle. The same node working for a long time might lead to temporary fluctuation of application resource demands and system hangs. In this situation, the load balancer will define a threshold value and regularly checks that the skewness values of all the nodes does not exceed this threshold value. By following this procedure make sure that the node being utilized for a long time should be freed and the workload is to be passed on to the nodes that are idle for a long time. If the work load is minimum and some nodes are idle then those nodes can be turned off temporarily thereby saving the energy. Thus , by reducing the temperature and saving the energy green computing is achieved.

Preethi [11] has discussed about Least Virtual Machine Assign algorithm to optimize the resources in cloud computing. This system shows the way for the green computing by allocating the virtual machine based on the load it is processing for the optimization of number of servers in use. The performance of the algorithm is analyzed using CloudSim simulator .The simulation result ensures that all the processors in the system as well as in the network does approximately equal amount of work at any instant of time by comparing with other algorithms.

A) *Least Virtual Machine(VM) Algorithm*:This algorithm is compared with Active VM load balancer algorithm. The main aim is to distribute the workload among available VM efficiently , so that the resources are not over or under-utilized. Initially , all the VMs are assigned to zero. If the VM is used already used , then its value is incremented. Then , the VM having least value is assigned the load. If the selected VM is not free , then it is excluded from the VM list.

B) *Cloudsim Simulator*: The main aim of simulator is to test the implementation work in the absence of the required environment. In the cloud environment two simulator are used CloudSim and Vcloud. CloudSim is the open source. It is a new generalized and extensible simulation framework that enables seamless modeling, simulation, experimentation of emerging cloud computing infrastructures and management services.

The simulation framework has the following novel features:

- It support for modeling and instantiation of large scale Cloud computing infrastructure, including data centers on a single physical computing node and java virtual machine.
- A self-contained platform for modeling data centers, service brokers, scheduling, and allocations policies.
- Availability of virtualization engine, which aids in creation and management of multiple, independent, and co-hosted virtualized services on a data center node.
- Flexibility to switch between space-shared and time-shared allocation of processing cores to virtualized services.

Ashish Kumar Singh [12] has discussed about Scheduling Algorithm with load balancing in cloud computing environment. This addresses the issue and propose an algorithm for private cloud which has high throughput and for public cloud which address the issue of environment consciousness also with performance. To improve the throughput in private cloud Shortest Job First (SJF) is used for scheduling and to overcome from the problem of starvation. For load balancing monitor the load and dispatch the job to the least loaded Virtual Machine (VM). To gain more benefit and to have opportunity for future enhancement in public cloud environment consciousness is the key factor and for better performance and load balancing also desired. While load balancing improve the performance, the environment consciousness increase the profit of cloud providers.

A) *Proposed Algorithm for Private cloud*

Database of VM is maintained at cloud manager which have 4 fields:

- VM id
- Num field indicate the number of process it can process.
- Number of processes currently allocated to the VM.
- Load percent on the VM on PCB an extra tag field is used for the bound waiting. Ready queue is maintained by cloud manager which contain the list of all arrived jobs.

B) *Proposed Algorithm for Public Cloud*

- Cloud computing is pay-as-you-use model. So number of job rejected must be less so scheduling must consider this issue.
- Cloud provider has to gain more profit so scheduling algorithm must increase the profit of cloud provider.
- Algorithm must be environment conscious because Carbon emission rate of ICT (Information and Communication Technology) industry is now approximately equal to the aviation industry. So government impose carbon emission limit. If scheduling algorithm not cover this issue then cloud provider are not able expand their infrastructure.
- Environment conscious also provide the benefit of first two points. Because if cloud provide has more infrastructure then number of job rejection is also low and to meet the more demand of cloud user cloud provider has to increase infrastructure which in turn also increase their profit.
- Load balancing must be there so process must be migrated form over loaded VM to less loaded VM within data Centre. This thing will not be done continuously means it will be done on periodic base. So it will not increase more overhead. Cloud manager periodically monitors the status of the VMs for the distribution of the load, if an overloaded VM is found then the cloud manage migrates the load of the overloaded VM to the underutilized VM.

DivyaRastogi [13] has discussed about Ant Colony Optimization (ACO) in cloud computing. It is a new heuristic algorithm for the solving the combinatorial optimization problems. ACO is inspired from the ant colonies that work together in foraging behaviour shows that the ant has the intelligence system to find an optimal path from nest to source of food. On the way of searching, ants act as the expert agents and in the direction of movement they lay some pheromone on the ground, while an isolated ant runs into a previously set trail. The trailing ant can detect it and decide to follow it with high probability. The probability of ant chooses a way is proportion to the concentration of a pheromone. The more ants chooses a way, which has denser pheromone, and vice versa. By this feedback technique, ant can find out an optimal way finally. The ants work together in search of new sources of food and simultaneously use the existing food sources to shift the food back to the nest.

The author describes an algorithm in order to make the system more reliable and fault tolerant. As cloud computing is a distributed system. For a distributed system to work properly, it is necessary that the system should be in safe state. In case of any fault, there should be technique to handle this fault so that working of the system should not get affect.

So here, proposed a model for fault management. As discussed above, ants follow the route according to the updated entries in pheromone table so it is necessary to store the status of this pheromone trail. To make the ACO more reliable with the capacity of fault tolerance, construct the system with the following features:

- Each ant will be having a small memory element.
- Every node in the system will have the knowledge about its neighbours.
- Each node can flood the message in the network.

Memory element is installed in each ant at the time of the construction of ACO system and this will work as knowledge base. Memory element will store the information about each movement of the ant in the system, its neighbouring ant's information and routing information. Knowledge base is used when a fault is generated and helps to take the corrective actions.

III . CONCLUSION AND FUTURE WORK

The load balancing is one of the greatest issue in Cloud Computing environment. To solve this issue, various techniques are used by researchers in the past . In this paper, we have discussed various existing load balancing methods in cloud computing environment. Load balancing is the requirement of a cloud environment and how well this requirement is met depends on the algorithm chosen. The various load balancing techniques are also being compared . The performance of the load balancing algorithms is evaluated by different parameters like throughput, response time, execution time, total cost and so on. Other parameters like fault tolerance can be included in the future for the better utilization and needs of the user.

REFERENCES

[1] Amritpal Singh,"*A Review of Existing Load Balancing Techniques in Cloud Computing*", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),Volume 4 Issue 7, July 2015.*

[2]Anjali, Jitender Grover, Anjali, Jitender Grover "*A New Approach for Dynamic Load Balancing in CloudComputing", IOSR Journal of Computer Engineering-ISSN: 2278-0661,p-ISSN: 2278-8727,PP 30-36,2015.*

[3] Reena Panwar , Prof. Dr. Bhawna Mallick *"Load Balancing in Cloud Computing Using Dynamic Load Management Algorithm"*Interntional Conference on Green Computing and Internet of Things,IEEE 978-1-4673-7910-6/15/$31.00 ©2015 IEEE, 2015.

[4] Surbhi Kapoor ,Dr. Chetna Dabas ,"*Cluster Based Load Balancing in Cloud Computing*" ,978-1-4673-7948-9/15/$31.00 ©2015 IEEE, 2015.

[5] Tejinder Sharma, Vijay Kumar Banga ,"*Efficient and Enhanced Algorithm in Cloud Computing"* ,International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-1, 2013.

[6] Priyank Singhal, Sumiran Shah, Sumiran Shah *"Load Balancing Algorithm over a Distributed Cloud Network".*

[7] Rajesh George Rajan, V.Jeyakrishnan, "*A Survey on Load Balancing in Cloud Computing Environments"* ,International Journal of Advanced Research in Computer and Communication EngineeringVol. 2, Issue 12, December 2013.

[8]. Suriya Begum, Prashanth CSR "*Mathematical Modelling of Joint Routing and Scheduling for an Effective Load Balancing in Cloud" ,International Journal of Computer Applications, Volume 104 – No.4, October 2014.*

9] Neha Gupta, Parminder Singh, "*Load Balancing Using Genetic Algorithm in Mobile Cloud Computing*", IJIET, Vol.4 Issue 1 June 2014.

[10] B. Bhaskar, E. Madhusudhana Reddy,"*A Novel Load Balancing Model Using RR Algorithm for the Cloud Computing*", IJCSIT, Vol. 5(6), 2014.

[11] B Preethi, Prof C. Kamalanathan, Dr. S.M Ramesh, S Shanmathi, P SathiyaBama, "*Optimization Of Resources in Cloud Computing Using Effective Load Balancing Algorithms*", IARJSET, Vol. 1, Issue 1, September 2014.

[12] Ashish Kumar Singh, Sandeep Sahu, MangalNath Tiwari, R.K katare, "*Scheduling Algorithm with Load Balancing in Cloud Computing*", IJSER, Vol. 2, Issue 1, Jan 2014.

[13] DivyaRastogi, FarhatUllah Khan, "*Effective Fault Handling Algorithm For Load Balancing Using Ant Colony Optimization in Cloud Computing*", IJAET, Vol. 7, Issue 3, july 2014.

[14]https://en.wikipedia.org/wiki/cloud_computing

.

[15]www.thbs.com/downloads/Cloud-Computing-Overview.pdf.

[16] Rajesh George Rajan, V. Jeyakrishnan, "A Survey on Load Balancing in Cloud Computing Environments", IJARCCE, Vol. 2, Issue 12, Dec 2013.

[17] Tushar Desai, JigneshPrajapati, "*A Survey of various load balancing techniques and challenges in cloud computing*", IJSTR, Vol. 2, Issue 11, Nov 2013.

[18] Pranjali D. Dhore, Dr. Kishor R. Kolhe, "*A Survey for Dynamic Balancing of workload and Scalability in Cloud Environment*", IJAPRR, Vol. 2, Issue 1, 2015.

Other Publications :

[9] Suriya Begum ,Prashanth ―Review of Load Balancing in Cloud Computing‖ ,IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 2, January 2013 ISSN (Print): 1694-0784 | ISSN (Online): 1694-0814

[10] Suriya Begum,Prashanth ―Investigational Study of 7 Effective Schemes of Load Balancing in Cloud Computing ―,IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 1, November 2013 ,ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784

. [11] Suriya Begum, Dr. Prashanth C.S.R,―Mathematical Modelling of Joint Routing and Scheduling for an Effective Load Balancing in Cloud‖ in International Journal of Computer Applications (0975 –8887) Volume 104 –No.4, October 2014.