Institute of Industrial and Computer Management and Research , Nigdi Pune.
MCA – 1st (Sem. II)
(2022 - 2023)

An Alternative Approach to Resolve Load Balancing Problem in Cloud Computing

Mansi Kishor Sisodiya,
Institute of Industrial and Computer Management and Research , Nigdi Pune
mansisisodiya223@gmail.com
8600539561

Mayur Deepak Bhavsar,
Institute of Industrial and Computer Management and Research , Nigdi Pune
Bhavsarmayur341@gmail.com
9372637531

Mokshada Bhandare,
Institute of Industrial and Computer Management and Research , Nigdi Pune
mokshadabhandare.01@gmail.com
7057171160

An Alternative Approach to Resolve Load Balancing Problem in Cloud Computing.

***Abstract*—**
Cloud computing has a decentralized architecture in which virtual machine migration is one of the major challenges which affects the network performance. To balance the network load, various techniques are designed for the virtual machine migration. In the previous research work, genetic algorithm was proposed for Virtual Machine (VM) migration which can balance the network load. The genetic algorithm is complex in nature which increases the execution time. In this research work, genetic algorithm is improved for VM migration which reduces the execution time and also space and bandwidth utilization.

***KEYWORDS*—** *Genetic Algorithm, Virtual Machine migration, Bandwidth utilization*

I. INTRODUCTION

Cloud Computing is service which provides on-demand and simple access of the network to several servers which provides computing resources like applications, storage, networks are in presence for other services which cloud provides which we can use to gain maximum efficiency. User retrieved data and modified data which is stored by client or an organization in centralized data called cloud. Cloud is a design, where cloud service provider provides services to user on demand and this vital feature is known as CSP stands for "Cloud Service Provider". Cloud computing gives Shared pool of assets (computers resources like networks, server and storage) on the demand of the user in ubiquitous and in simple word provides usability to the end customer maintaining minimum cost for their usage. According to NIST definition the basic concept of cloud computing can be easily understood [1]. From the above definition, we can conclude that cloud computing provides virtual infrastructure backed by software and hardware resources available on the internet. Cloud computing gives access to a user's shared pool of resources on demand of the user, to which the user subscribe and use for the time he wants to use and this all is achieved with the help of virtualization, which further helps in reducing the cost of implementing or adding more hardware parts to achieve the requirements of the user [2]. In cloud computing, there is no need to know the physical location, configuration of the system which provides the service. The basic features of cloud computing are: virtualization, homogeneity, modern security, on demand scale, minimum cost software, geographic distribution service orientation. One can use the application without installation and just by accessing internet user can manage

II. LOAD BALANCING STRATEGIES FOR CLOUD.

Load balancing algorithms can be broadly categorized into static and dynamic load balancing algorithms.

Static load balancing algorithms: Gulati et al. [24] claimed that in cloud environment a lot of work is done on load balancing in homogeneous resources. Research on load balancing in heterogeneous environment is given also under spot light. They studied the effect of round robin technique with

dynamic approach by varying host bandwidth, cloudlet long length, VM image size and VM bandwidth. Load is optimized by varying these parameters. CloudSim is used for this implementation.

Dynamic load balancing algorithms: A hybrid load balancing policy was presented by Shu-Ching et al. [25]. This policy comprises of two stages 1) Static load balancing stage 2) Dynamic load balancing stage. It selects suitable node set in the static load balancing stage and keeps a balance of tasks and resources in dynamic load balancing stage. When a request arrives a dispatcher sends out an agent that gathers nodes information like remaining CPU capacity and memory. Hence the duty of the dispatcher is not only to monitor and select effective nodes but also to assign tasks to the nodes accordingly. Their results showed that this policy can provide better results in comparison with min-min and minimum completion time (MCT), in terms of overall performance.

Another algorithm for load balancing in cloud environment is ant colony optimization (ACO) [26]. This work basically proposed a modified version of ACO. Ants move in forward and backward directions in order to keep track of overloaded and under loaded nodes. While doing so ants update the pheromone, which keeps the nodes' resource information. The two types of pheromone updates are 1) Foraging pheromone, which is looked up when an under loaded node is encountered in order to look for the path to an over loaded node. 2) Trailing pheromone is used to find path towards an under loaded node when an over loaded node is encountered. In the previous algorithm ants maintained their own result sets and were combined at a later stage but in this version these result sets are continuously updated. This modification helps this algorithm perform better.

Genetic algorithm [27] is also a nature inspired algorithm. It is modified by Pop et al. [28], to make it a reputation guided algorithm. They evaluated their solution by taking load-balancing as a way to calculate the optimization offered to providers and makespan as a performance metric for the user.

Another such algorithm is the bees life algorithm (BLA) [29], which is inspired by bee's food searching and reproduction. This concept is further extended to specifically address the issue of load balancing in [30]. The Honey bee behavior inspired load balancing (HBBLB) algorithm basically manages the load across different virtual machines for increasing throughput. Tasks are prioritized so that the waiting time is reduced when they are aligned in queues. The honey bee foraging behavior and some of its variants are listed in [31]

III. LOAD BALANCING STRATEGIES IN HYBRID CLOUD.

Zhang et al. [36] proposed a design for hybrid cloud is. It allows intelligent workload factoring by dividing it into base and trespassing load. When a system goes into a panic mode the excess load is passed on to the trespassing zone. Fast frequent data item detection algorithm is used for this purpose. It makes use of the least connections balancing algorithm and the Round-Robin balancing algorithm as well. Their results show that there is a decrease in annual bills when hybrid clouds are used. Buyya et al. [37] proposed a concept of federated cloud environment, to maintain the promised QoS even when the load shows a sudden variation. It supports dynamic allocation of VMs, Database, Services and Storage. That allows an application to run on clouds from different vendors. In Social Networks like Facebook, load varies significantly from time to time. For such systems this facility can help scale the load dynamically. No cloud infrastructure provider can have data centers all around the globe. That's why to meet QoS, any cloud

application service provider has to make use of multiple cloud providers. For implementation purpose they used Cloud Sim Tool kit. They made a comparison between federated and non federated cloud environments. Their results showed a considerable gain in performance in terms of response time and cost in case of the former. The turnaround time is reduced by 50% and the make span improves by 20%. Although the overall cost increases with the increase in the public cloud utilization but one has to consider that such peak loads are faced occasionally which makes it acceptable.

Task scheduling plays a vital role in solving the optimization problem in hybrid clouds. A graph-based task scheduling algorithm is proposed by Jiang et al. for this purpose [38]. In order to reduce the cost to a minimum value, like other algorithms, it makes use of the public resources along with the private infrastructure. The key stages of this algorithm are 1) Resource discovery and filtering, for the collection of the status information of the resources that are discovered. 2) Resource selection, this algorithm's main focus is on this stage as this is the decision making stage. Resources are picked keeping in view the demand of the tasks to be performed. 3) Task submission, once the resources are selected the tasks are assigned accordingly. A bipartite graph G=(U,V,E) is used to help elaborate this concept, where U is used for private or public Virtual Machines, V is for the tasks, and E denotes the edges in between. Cloud Report and Cloud Sim 3.0 are used for evaluating this algorithm. Their results showed a 30 % decrease in cost as compared to a non hybrid environment. For improving these figures even more, disk storage and network bandwidth need to be considered as well.

Another algorithm, adaptive-scheduling-with-QoS-satisfaction (AsQ) [39], for the hybrid cloud is proposed that basically reduces the response time and helps increase the resource utilization. To fulfill this goal several fast Scheduling strategies and run time estimations are used and resources are then allocated accordingly. If resources are used optimally in the private clouds, the need for transferring tasks to the public clouds decreases and deadlines are fulfilled efficiently but if a task is transferred to the public cloud, minimum cost strategy is used so that the cost of using a public cloud can be reduced. The size of the workload is specially considered in this regard. Their results show that As Q performs better compared to the recent algorithms of similar nature in terms of task waiting, execution and finish time. Hence it provides better QoS.

Picking the best resources from the public cloud is a serious concern in hybrid clouds. The Hybrid Cloud Optimized Cost (HCOC) [40], is one such scheduling algorithm. It helps in executing a workflow within the desired execution time. Their results have shown that it reduces the cost while meeting the desired goals. Gives better results in comparison with the other greedy approaches. There is another approach [41], which also deals with directed acyclic graphs (DAG) as in study by Bittencourt and Madeira [40]. It uses integer linear program (ILP) for the workflow scheduling n SaaS/PaaS clouds with two levels of SLA, one with the customer one for the provider. This work can be extended by considering multiple workflows and fault tolerance in view.

Gupta et al. [42], contributed that there are a number of load balancing algorithms that basically help in avoiding situations where a single node is loaded heavily and the rest are either idle or have lesser number of tasks when in reality they can afford to deal with a lot more. But what is overlooked in most of these algorithms is the trust and reliability of the data center. A suitable trust model and a load balancing algorithm are proposed. They used VMMs (Virtual Machines Monitors) to generate trust values

on the basis of these values nodes are selected and the load is balanced.

A virtual infrastructure management tool is offered by Hoecke et al. [43], that helps to set-up and manage hybrid clouds in an efficient way. This tool automatically balances load between the private and public clouds. It works at the virtual machine level. This tool has two parts 1) a proxy, where different load balancing algorithms are implemented like weighted round robin and forwarding requests to appropriate VMs, and on the other hand a management interface is designed that visualizes the hybrid environment and manages it too for example it can start and stop VMs, can form clusters of VMs, and can also manage the proxy remotely. It can be improved further by using a more efficient algorithm on the proxy for balancing load in a more convenient way.

In workflow applications [44], the cost of execution is kept to a minimum level by allocating the workflow to a private cloud but in case of peak loads, resources from the Public cloud need to be considered as well. As meeting the deadlines is a primary concern in workflow applications. By using cost optimization, this algorithm decides which resources should be leased from the public cloud for executing the task within the deadline. In this algorithm workflow is divided into levels and scheduling is performed on each level. It uses the concept of subdeadlines as well. That helps in finding the best resources in public cloud in terms of cost while keeping in view that the workflows are executed within the deadlines. Although the make span of level based approach is 1.55 times higher than the non level based approach, its cost is three times lower. In comparison with min-min, its make span is double but it costs three times lesser. This makes the proposed level based approach better as it costs less and meets the deadlines too although its make span is higher but it finishes the assigned tasks within the deadline.

IV. MATRICE OF LOAD BALANCING:

The different qualitative metrics or parameters that. are considered important for load balancing in cloud computing [18] are discussed as follows:

1. **Throughput**: The total number of tasks that have completed execution is called throughput. A high throughput is required for better performance of the system.
2. **Associated Overhead**: The amount of overhead that is produced by the execution of the load balancing algorithm. Minimum overhead is expected for successful implementation of the algorithm.
3. **Fault tolerant:** It is the ability of the algorithm to perform correctly and uniformly even in conditions of failure at any arbitrary node in the system.
4. **Migration time:** The time taken in migration or transfer of a task from one machine to any other machine in the system. This time should be minimum for improving the performance of the system.
5. **Response time**: It is the minimum time that a distributed system executing a specific load balancing algorithm takes to respond.
6. **Resource Utilization:** It is the degree to which the resources of the system are utilized. A good load balancing algorithm provides maximum resource utilization.
7. **Scalability:** It determines the ability of the system to accomplish load balancing algorithm with a restricted number of processors or machines.
8. **Performance:** It represents the effectiveness of the system after performing load balancing. If all the above parameters are satisfied optimally then it will highly improve the performance of the system.

V. CONCLUSION:

Load balancing is one of the main challenges in cloud computing. It is required to distribute the load evenly at every node. A highly congested provider may fall to provide efficient services to its customers. So, with proper load balancing algorithm system service and throughput can be increased. This paper is to focus on one of the major concerns of cloud computing that is Load balancing. The goal of load balancing is to increase client satisfaction and maximize resource utilization and substantially increase the performance of the cloud system thereby reducing the energy consumed and the carbon emission rate.

Load balancing algorithms can be broadly categorized into static and dynamic load balancing algorithms. A comparative study of different load balancing algorithms is presented. Load balancing is not only required for meeting users' satisfaction but it also helps in proper utilization of the resources available. The metrics that are used for evaluating different load balancing technologies are: throughput, overhead associated, fault tolerance, migration time, response time, resource utilization, scalability, and performance. According to this study, in honeybee foraging algorithm, throughput does not increase with the increase in system size. Different load balancing strategies in hybrid cloud are also discussed. A concept of federated cloud environment is proposed to maintain the promised QoS even when the load shows a sudden variation. Task scheduling plays a vital role in solving the optimization Problem in hybrid clouds. Another algorithm, adaptive-scheduling-with-QoS-satisfaction (AsQ) for the hybrid cloud is proposed that basically reduces the response time and helps increase the resource utilization.

VI. PROPOSED MODEL:

The proposed load balancing model will use improved throttle algorithm. This improved throttle algorithm works well even though underlying capacity of each VM is different because the hardware configuration of VMs is different. So improved throttle algorithm is taking decision of VM selection with hash table with more parameters such as expected response time and loading condition. Now Expected response time can be calculated using CPU utilization of VM. Using modified throttled load balancing algorithm with less overhead, results better VM allocation and increased number of user request handling, thus reducing the rejection in the number of requests arrived at datacenter of cloud.

VII. PROPOSED SYSTEM MODEL:

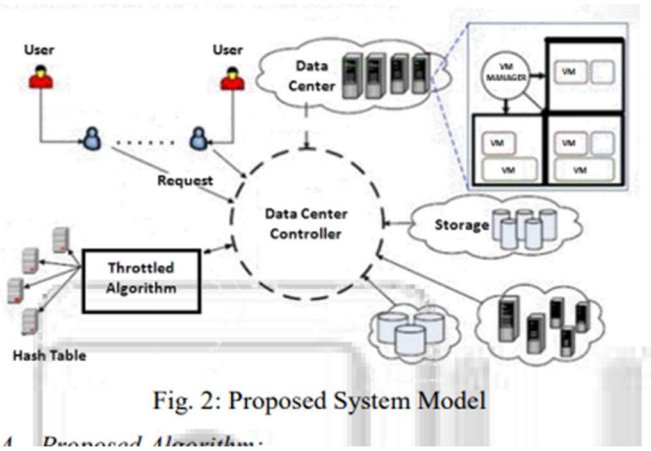

Fig. 2: Proposed System Model

A. Proposed Algorithm:

- **Input:**

Data centre requests r1,r2,….., rn

Available VMs vm1,vm2, ,vmn

- **Output:**

Data centre requests r1,r2,……,rn are allocated available

VMs vm1,vm2,…….,vmn

Steps:

1) The improved throttled algorithm maintains a hash map table of all the available VMs which their current state and the expected response time. This state may be available or busy. At the beginning, all the VMs are available.

2) When data centre controller receives a request the it forwards that request to the improved throttled load balancer. The improved throttled load balancer is responsible for the VM allocation. So that the job can be accomplished.

3) The improved throttled algorithm scans the hash map table. It checks the status of the available VMs.If a VM with least load and the minimum response time is found.

- Then the improved throttled algorithm sends the VM id of that machine to the data centre controller
- Data centre controller sends a request to that VM
- Data centre controller sends a notification of this new allocation to the improved throttled
- The improved throttled algorithm updates the hash map index accordingly.

If a VM is not found then the improved throttled algorithm returns -1 to the data centre controller

4) When the VM finishes the request. – The data centre controller sends a notification to improve throttled that the VM id has finished the request. – improved throttled modifies the hash map table accordingly

5) If there are more requests then the data centre controller repeats step 3 for other VMs until the size of the hash map table is reached. Also of the size of hash map table is reached then the parsing starts with the first hash map index.

VIII. REFERENCES:

[1] Srinivas.J, K. Venkata Subba Reddy, Dr. A. Moiz Qyser, "Cloud Computing Basics", International journal of advanced research in computer and communication engineering , 2012, pp. 343-347.

[2] Soumya Ray and Ajanta De Sarkar, "Execution Analysis of Load Balancing Algorithm in Cloud computing Environment", International Journal on Cloud Computing: Services and Architecture (IJCCSA), October 2012, Vol.2, No.5.

[3] HU Baofang, SUN Xiuli, LI Ying, SUN Hongfeng, "An Improved Adaptive Genetic Algorithm in Cloud Computing", 13th International Conference on Parallel and Distributed Computing, Applications and Technologies, 2012.

[4] Tushar Desai, Jignesh Prajapati, "A Survey of Various Load Balancing Techniques and Challenges in Cloud Computing", International Journal of Scientific & Technology Research, November 2013, Volume 2, Issue 11.

[5] Punithasurya K, Esther Daniel, Dr. N. A. Vasanthi, "A Novel Role Based Cross Domain Access Control Scheme for Cloud Storage", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2013, Volume 2, Issue 3, March 2013, pp 942-946.

[6] Vimmi Pandey, "Securing the Cloud Environment Using OTP", International Journal of Scientific Research in Computer Science and Engineering, 2013, vol-1, Issue-4.

[7] Sanjoli Singla, Jasmeet Singh, "Cloud Data Security using Authentication and Encryption Technique", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) , July 2013 Volume 2, Issue 7, pp 2232-2235.

[8] Sheetal Karki, Anshika Goyal, "Performance Evaluation of Check Pointing and Threshold Algorithm for Load Balancing in Cloud Computing", International Journal of Computer Sciences and Engineering, Vol.-6, Issue-5, May 2018, pp 2347-2693.

[9] Sukhpreet Kaur, Dr. Jyotsna Sengupta, "Load Balancing using Improved Genetic Algorithm(IGA) in Cloud Computing", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 6, Issue 8, August 2017, pp 2278-1323.

[10] Wang Bei, LI Jun, "Load Balancing Task Scheduling based on Multi-Population Genetic Algorithm in Cloud Computing", 2016, Proceedings of the 35th Chinese Control Conference

[11] Mahalingam, Nandhalakshmi Nithya, "Efficient Load Balancing in Cloud Computing Using Weighted Throttled Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, 2015,vol.3, 5409 – 5415.

[12] Keke Gai, Meikang Qiu, Hui Zhao, "Cost-Aware Multimedia Data Allocation for Heterogeneous Memory Using Genetic Algorithm in Cloud Computing", IEEE Transactions on Cloud Computing, 2015.

[13] Mr. Mayur S. Pilavare, Mr. Amish Desai, "A Novel Approach Towards Improving Performance of Load Balancing Using Genetic Algorithm in Cloud Computing", IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems, ICIIECS'15, 2015.