EasyChair Preprint

№ 4914

A Novel approach on Load Balancing in Cloud Computing System

H R Manjunatha

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

January 19, 2021

A Novel approach on Load Balancing in Cloud Computing System

Manjunatha H R
Asst Professsor, Dept. of CSE
BGS Institute of Technology, BG Nagar
E-Mail: manjuhr19@gmail.com.

Abstract— **Cloud Computing is the pool of virtualized computer resources. Upcoming generation of cloud computing will boom in filed how to use the available infrastructure and resources effecti vel y by providing good quality of ser vices. Cloud computing is the Internet-based development where dynamically scalable and often vir tualized resources is provi ded as a ser vice to user over the Internet has become a major issue. To provide such service Load balancing which is one of the main challenges in Cloud computing, distributes the dynamic workload across multiple nodes to ensure that no single resource is either over whelmed or underutilized. This can be considered as an optimization problem and a good load balancer should adapt its strategy to the changing environment and the types oftasks.**

Keywords: - Cloud, Load Balancing, Virtual Machine

I. INTRODUCTION

Cloud computing is one of the leading technology in the IT sector. Computing in frastructure is used by businesses and individuals to access storage from remote place anywhere in the world on demand. Any cloud service provider offers computing, storage, and software "as a service". Cloud computing accommodates provisioning and de-provisioning on demand and helps any organization in avoiding the capital costs of software and hardware. Due to the exponential growth of cloud computing it has been widely adopted by the industry and thus making a rapid e xpansion in availability of resource in the Internet. As the size of cloud scales up cloud computing service providers requires handling of massive requests. The primary challenge then becomes to keep the performance same or better whenever such an outburst occurs. Thus in spite of glorious future of Cloud Computing, many critical problems still need to be e xp lored for its perfect realization. One of these issues is Load balancing.

II. LOAD BALANCING

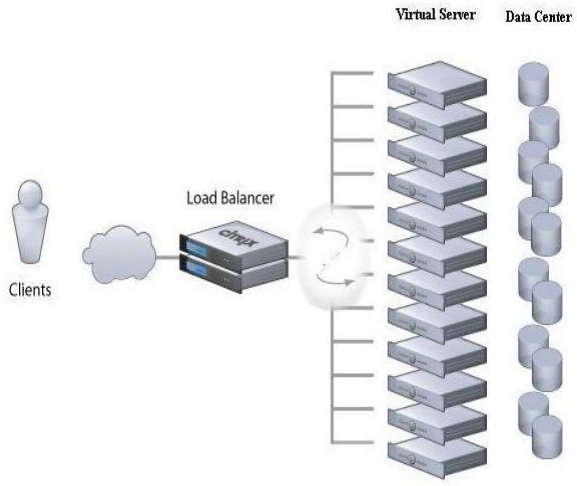
Load balancing is the process of reassigning the total loads to the individual nodes of the collective system to make the best response time and also good utilization of the resources. Cloud computing is an internet computing in which the load balancing is the one of the challenging task. Cloud Computing is made up by aggregating two terms in the field of technology. First term is Cloud and the second term is computing. Cloud is a pool of heterogeneous resources. It is a mesh of huge infrastructure and has no relevance with its name " Cloud". Infrastructure refers to both the applications delivered to end users as services over the Internet and the hardware and system software in datacenters that is responsible for providing those services. Various methods are to be used to make a better system by allocating the loads to the nodes in a balancing manner but due to network congestion, bandwidth usage etc., there were problems are occurred. These problems were solved by some of the existing techniques. A load balancing algorithm which is dynamic in nature does not consider the previous state or behavior of the system, that is, it depends on the current behavior of the system. There were various goals that related to the load balancing such as to improve the performance substantially, to maintain the system stability etc. Depending on the current state of the system, load balancing algorithms can be categorized into two types they are static and dynamic algorithms. static algorithm: In static algorithm the traffic is divided evenly among the servers. This algorithm requires a prior knowledge of system resources, so that the decision of shifting of the load does not depend on the current state of system. Static algorithm is proper in the system which has low variation in load .dynamic algorithm: In dynamic algorithm the lightest server in the whole network or system is searched and preferred for balancing a load. For this real time communication with network is needed which can increase the traffic in the system. Here current state of the system is used to make decisions to manage the load.

III. NEED FOR LOAD BALANCING IN CLOUD

Distribution of the inbound IP traffic across multiple servers is called load balancing. It increases the performance, leads to optimal utilization and ensures that no single server is overwhelmed. We can have multiple servers in a server farm or a data center, which can host multiple guests. Each guest may dire on load, leading to a situation where some of the servers may become overwhelmed or capitulated in terms of computational resources, memory resources and I/O devices. For simplicity and without loss of generality, we will consider load in terms of CPU t ime. When a server A is unable to allocate su_cient CPU time slice due to heavy demand by

other VMs running parallel and another server B is id le, we can redistribute the load from A to B by migrating few guests to B. This would be an ideal situation when we require load balancing policy in Cloud Computing scenario, improving the percentage of server idle times, marginal job response times, etc. In order to achieve, short response time and high system throughput, we need to consider the following characteristics:

- The load balancing process generates little trace overhead and adds low overhead on the computational and network resources.
- It keeps up to date load information of the participating systems.
- It balances the system uniformly and takes action instantaneously or on a periodic basis.
- It can run on a dedicated system, or it can be a decentralized.
- The available server should have sufficient resources available to host and run the migrated guest.
- The live migration of complete operating system should take minimum acceptable time with minimum downtime.
- Network communication should be reliable and fast.



IV. RESERCH ISSUES

- In local distribution system minimization of problem of allocating the considerable processing capacity to utilize the full advantage of processing capacity

.

- In cloud system reduction in the chance of outages so has to provide good quality of service without interruption
- Dividing the traffic between servers, data can be sent and received without major delay. Without load balancing, users could experience delays, timeouts and possible long system responses.
- Objective the minimization of overall e xpected response time. The fairness of allocation, which is an important issue for modern distributed systems.
- Control of Routing traffic in networks

V. EXISTING LOAD BALANCING TECHNIQUES

To balance the load following are the existing techniques

1) A fast adaptive load balancing method: a binary tree structure that is used to partition the simulation region into sub-domains. The characteristics of this fast adaptive balancing method are to be adjusted the workload between the processors from local areas to global areas. According to the difference of workload, the arrangements of the cells are obtained. But the main workload concentrates on certain cells so that the procedure of adjusting the vertices of the grid can be very long because of the local workload can be considered. This problem can be avoided by the fast load balancing adaptive method. Here the region should be partitioned by using the binary tree mode, so that it contains leaf nodes, child nodes, parent nodes etc. There were partition line between the Binary tree and the indexes of the cells on the left are smaller that of right and the indexes on the top are smaller than the bottom. Calculate the workload based on the balancing algorithm. This algorithm has a faster balancing speed, less elapsed time and less communication time cost of the simulation procedure. Advantages are Relative smaller communication overhead relative smaller communication overhead, faster balancing speed, and high efficiency and the disadvantage is it cannot maintain the topology that is neighboring cells cannot be maintained.

2) Honey Bee Behavior Inspired Load Balancing: an algorithm named honeybee behavior inspired load balancing algorithm. Here in this session well load balance across the virtual machines for maximizing the throughput. The load balancing cloud computing can be achieved by mode ling the foraging behavior of honey bees. This algorithm is derived from the behavior of honey bees that uses the method to find and reap food. In bee hives, there is a class of bees called the scout bees and the another type was forager bees .The scout bee which forage for food sources, when they find the food, they come back to the beehive to advertise this news by using a dance called waggle/tremble/vibration dance. The purpose of this dance, gives the idea of the quality and/or quantity of food and also its distance from the beehive. Forager bees then follow the Scout Bees to the location that they found food and then begin to reap it. After that they return to the beehive and do a tremble or vibration dance to other bees in the hive giving an idea of how much food is left. The tasks removed from the overloaded VMs act as Honey Bees. Upon submission to the under load VM , it will update the number of various priority tasks and load of tasks assigned to that VM . This information will be helpful for other tasks , i.e., whenever a high priority has to be submitted to VMs, it should consider the VM that has a minimum number of high priority tasks so that the particular task will be e xecuted earlier. Since all VMs are sorted in an ascending order, the task removed will be submitted to under loaded VMs. Current workload of all available VMs can be calculated based on the information received from the data center. Advantages are maximizing the throughput; waiting time on task is minimum and overhead

become minimum. The disadvantage is if more priority based queues are there then the lower priority load can be stay continuously in the queue.

3) A Dynamic and Adaptive Load Balancing Strategy For Parallel File System: a dynamic file migration load balancing algorithm based on distributed architecture. Considered the large file system there were various problems like dynamic file migration, algorithm based only on centralized system etc. So these problems are to be avoided by the introduction of the algorithm called self-acting load balancing algorithm (SALB).In the parallel file system the data are transferred between the memory and the storage devices so that the data management is an important role of the parallel file system. There were various challenges that are faced during load balancing in the parallel file system. They are scalability and the availability of the system, network transmission and the load migration. Considered the dynamic load balancing algorithms, the load in each I/ O servers are different because the workload becomes varies continuously. So there were some decision making algorithms are needed. In this decision making system, there were firstly central decision maker, by which the central node is the decision maker so that if the central node becomes fail, then the whole system performance becomes down and the reliability becomes less. Secondly group decision maker in which the total system should be divided in to groups so that the communication cost becomes reduced. But taking decision without considered the whole system load so that global optimization e xp lored a major problem. Finally the distributed decision maker in which each I/O server can take their own decision so that they provide better scalability and availability. This proposed SALB addressed the load prediction algorithm, efficient load collection mechanism, effective distributed decision maker, migration selection model and dynamic file migration algorithm for a better load balancing. The disadvantage is degradation of the whole system due to the migration side effect.

4) Heat Diffusion Based Dynamic Load Balancing: an efficient cell selection scheme and two heat diffusion based algorithm called global and local diffusion. Considered the distributed virtual environments there were various numbers of users and the load accessing by the concurrent users can cause problem. This can be avoided by this algorithm. According to the heat diffusion algorithm, the virtual environment is divided in to large number of square cells and each square cell having objects. The working of the heat diffusion algorithm is in such a way that every node in the cell sends load to its neighboring nodes in every iterat ion and the transfer was the difference between the current nodes to that of neighboring node. So it was related to heat diffusion process. That is the transfer of heat from high to low object, when they were placed adjacently in local diffusion algorithm, there were local decision making and efficient cell selection schemes are used. Here they simply compared the neighboring node loads to the adjacent node loads. If load is small then the transfer of load becomes possible. When global diffusion algorithm considered, it has two stages that is global scheduling stage and local load migration stage. From various experimental results the global diffusion algorithm becomes the better one. Advantages are communication overhead is less, high speed and require little amount of calculations. Disadvantages are network delay is high and several iterations are taken so there was a waste of time.

2.5Decentralized Scale-Free Network Construction and Load Balancing in Massive Multiuser Virtual Environments: the concept of overlay networks for the interconnection of machines that makes the backbone of an online environment. Virtual online world that makes the opportunities to the world for better technological advancements and developments. So the proposed network that makes better feasibility and load balancing to the dynamic virtual environments. This proposed system developed Hyper verse architecture, that can be responsible for the proper hosting of the virtual world. There were self-organized load balancing method by which the world surface is subdivided in to small cells, and it is managed by a public server. In this cells various hotspots so that the absolute mass of the object in the cell can be calculated by the public server. Hotspot accuracy is better when increasing the network load. The proposed algorithm cannot avoid the overloaded nodes but find out the number of links that assigned to each node while join ing the network. The advantages are the network becomes reliable; the network becomes resilience, efficient routing, and fault tolerant. The disadvantage is the overload ratio at the beginning is higher so that public servers are init ially placed randomly so some time is used for balancing the load.

6) Load Balancing in Dynamic Structured P2P Systems: an algorithm for load balancing in dynamic peer-to-peer system and other hybrid environments. In most peer-to-peer system the non-uniform of objects in the space and also the load of the node can be changed continuously due to the insertion, deletion and other various operations. This will leads to decrease the performance of the system. So the concept of virtual server can be introduced. In this proposed load balancing algorithm, the load information of the peer nodes are stored in different directories. These directories help to schedule reassignment of the virtual servers to develop a better balance. Greedy heuristic algorithm used to find out a better solution for the proper utilization of the nodes. The huge number of virtual servers in the system helps to increase the utilization. The various load information in to the corresponding pool and then the virtual server assignments are to be done. This proposed algorithm should be applied to different types of resources like storage, bandwidth etc., It was designed to handle the various situations like varying load of the node, node capacity, entering and leaving of nodes and also insertion and deletion of the nodes. Advantages are high node utilization and increasing scalability. Disadvantage is the reassignment of the virtual server is difficult.

7) Throttled Load Balancing Algorithm: Throttled algorithm is completely based on virtual machine. In this client first requesting the load balancer to check the right virtual machine which access that load easily and perform the operations which is given by the client or user. This ensures that only a

pre-defined number of internet cloud-lets are allocated to a single VM at any given time. If more request groups are present than the number of available VMs at a data center, some of the request will be queued until the next VM becomes available Throttled algorithm is completely based on virtual machine. In this client first requesting the load balancer to check the right virtual machine which access that load easily and perform the operations which is given by the client or user

8) Equally Spread Current: Execution Equally load distributing improves performance by transferring load from heavily loaded server. Efficient scheduling and resource allocation is a critical characteristic of cloud computing based on which the performance of the system is estimated. In spread spectrum technique load balancer makes effort to preserve equal load to all the VMs connected with the data center. Load balancer maintains an index table of VMs as well as number of requests currently assign to the VM. If the request comes from the data center to allocate a new VM, it scans the index table for the least loaded VM

9) Round Robin: It is the simplest algorithm that uses the concept of time quantum or slices. Here the time is divided into multiple slices and each node is given a particular time quantum or time interval and in this quantum the node will perform its operations. The resources of the service provider are provided to the client on the basis of this time quantum.

VI. Related Work

This section presents some of the prior research work that has been addressing the load balancing technique to provide quality service to user with the available resource.

Grosu et al. [1] proposed a cooperative load balancing game and present the structure of the NBS. For this game an algorithm for computing NBS is derived. They show that the fairness index is always 1 using NBS which means that the allocation is fair to all jobs. Finally, the performance of their cooperative load balancing scheme is compared with that of other existing schemes. Pathan and Mallikarjuna [2] presented a better load balance model for the Job Seeker's Web Portal based on the cloud partitioning concept with a switch mechanism to choose different strategies for different situations. Thus, this model divides the public cloud into several cloud partitions. When the environment is very large and comple x, these divisions simplify the load balancing. The cloud has a main controller that chooses the suitable partitions for arriving jobs based on arrival date. Thus with cloud partitioning concept it is possible to provide good load balancing and hence improving the overall performance of cloud environment and user satisfaction. Lanjewar et al. [3] illustrated days cloud computing is one of the greatest platform which provides storage of data in very lower cost and available for all t ime over the internet. But it has more critical issue like security, load management and fault tolerance. In this paper they are discussing Load Balancing approach. Many types of load concern with cloud like memory load, CPU load and network load. Load balancing is the process of distributing load over the different nodes which provides good resource utilization when nodes are overloaded with job. Load balancing has to handle the load when one node is overloaded. When node is overloaded at that time load is distributed over the other ideal nodes. Khare and Chauhan [4] presented various load balancing techniques for cloud partitioning. Load balancing in the cloud computing surroundings has an imperative impact on the performance. Excellent load balancing makes cloud computing more efficient and improves user satisfaction. More et al. [5] introduced a better load balance model for public cloud based on the cloud partitioning concept with a switch mechanism to choose different strategies for different situations. The algorithm applies the game theory for load balancing strategy to improve the efficiency in the public cloud environment. Khan et al. [6] proposed the new architecture for load balancing as well as performance enhancement in public cloud. The concept of load balancing is implemented on the basis of cloud partitioning method with different strategies for different cloud status. Their primary aim is to achieve the results and compare those with the existing system. Significant level of development is running to satisfy the required objectives. Sonawane et al. [7] have illustrated a cloud computing environment; load balancing has an important impact on the performance. Effective implementation of load balancing can make cloud computing more effective and it also improves user satisfaction. A better load balance model can be implemented for the large cloud which uses the cloud partitioning concept. Switch mechanism can also be used to choose different strategies for different situations. The algorithm applies the token generation algorithm to improve the effect of load balancing strategy in the public cloud environment. Singh and Phogat [8] demonstrated a cloud partitioning based dynamic load balancing strategy has been proposed for public cloud infrastructure. A Nash-equilibrium based non-cooperative game theoretic approach has been developed for heterogeneous distributed cloud system. To ensure data security on cloud a RSA cryptosystem based user authentication scheme has been implemented that ensures genuine resource utilization and secured data access on PCIs. The developed system has exhibited better for task scheduling and load balancing in PCI systems. Karthika et al. [9] developed an adaptive fault-tolerant quality of service control methods which is based on hop-by-hop data delivery utilizing "source" and "path" redundancy, with the goal to satisfy application of QoS requirements which prolongs the lifetime of the sensor system is explained. Thus their project also explains the better architecture of the cloud switching in different mechanism and also effectively balances the load with QOS and query aggregation process. Palivela et al. [10] presented an improved load balance model for the public cloud centered on the cloud segregating concept with a switch mechanism to select different approaches for different circumstances. The algorithm relates the game theory to the load balancing approach to increase the proficiency in the public cloud environment.

VII. CONCLUSION

The load balancing of the current system is one of the greatest issues. Various techniques and algorithms are used to solve the problem. Good load balance will improve the performance of the entire cloud. However, there is no common method that can adapt to all possible different situations. Various methods have been developed in improving existing solutions to resolve new problems. Each particular method has advantage in a particular area but not in all situations. In some, model integrates several methods and switches between the load balance methods based on the system status.

VIII. REFERENCES

[1] D. Grosu, A. T. Chronopoulos, and M. Y. Leung, Load balancing in distributed systems: An approach using cooperative games, in Proc. 16th IEEE Intl. Parallel and Distributed Processing Symp., Florida, USA, pp. 52-61, 2002

[2] A.F Pathan, S. B. Mallikarjuna, "A Load Balancing Model Based on Cloud Partitioning for the Public Cloud", International Journal of Information & Computation Technology, Vol.4, No. 16, pp. 1605-1610, 2014

[3] S. M. Lanjewar, S. S. Surwade,S. P. Patil,P. S. Ghumatkar, Y.B. Gurav, "Load Balancing In Public Cloud", IOSR Journal of Computer Engineering, Vo l. 16, Issue. 1, pp. 82-87, 2014

[4] S. Khare and A. Chauhan, "A Review on Load Balancing Model Based on Cloud Partitioning for the Public Cloud" International Journal of Emerging Technology and Advanced Engineering, Vo l. 4, Issue. 7, 2014

[5] S.D.More, S. Chaudhari, " Reviews of Load Balancing Based on Partitioning in Cloud Computing", International Journal of Computer Science and Information Technologies, Vo l. 5 (3) , pp.3965-3967, 2014

[6] N.G. Khan and V. B. Bhagat, "Cloud Partitioning Based Load Balancing Model for Performance Enhancement in Public Cloud", International Journal of Science and Research (IJSR) , ISSN. 2319-7064, 2012

[7] S.D. Sonawane and R.H.Borhade, " Load Distribution and Balancing over Cloud using Cloud partitioning", International Journal of Current Engineering and Technology, Vol. 4, No. 3, 2014

[8] Amritpal Singh and 2Nisha Phogat, "Cloud Partitioning Based Secured SecuredLoad balancing Approach alancing Approach alancing Approach for Public or Public Cloud Infrastructure Cloud Infrastructure", International Journal of Research in Engineering & Advanced Technology, Vo lume 2, Issue 2, Apr-May, 2014

[9] S.Karthika, T.Lavanya, 3G.Go kila, 4A.Arunraja 5 S.Sarumathi, 6 S.Saravanakumar, 7A.Go kilavani, " Load Balancing and Maintaining the Qos on Cloud Partitioning For the Public Cloud", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue. 2, 2014

[10] R. K. Palivela, C. S. Redy, "OAD Balancer Strategy Based On Cloud Computing", ernational Journal of Research in Computer and Communication Technology, Vol 3, Isue 10, October - 2014

[11] Gaochao Xu, Junjie Pang, and Xiaodong Fu, A Load Balancing Model Based on Cloud Partitioning for the Public Cloud, IEEE transactions on cloud computing , 2013