Volume 6, Issue 4, April 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey of Existing Load Balancing Algorithms in Cloud Computing

Mohit
M.Tech CSE Scholar, Department of Computer Science and Engineering, DCRUST, Murthal, Haryana, India

Jitender Kumar
Assistant Professor, Department of Computer Science and Engineering, DCRUST, Haryana, Murthal, India

Abstract– In today's era as IT industry is growing day by day the need of computing and storage is increasing rapidly. Users are demanding for services with better results. For this cloud computing requires efficient load balancing techniques. Load balancing is essential for efficient operations in distributed environments. Load balancing is a way of proper utilization of resources. Many algorithms were suggested to provide various approaches for the client's request to cloud nodes. There algorithms are used to enhance the performance of cloud with the aim of user satisfaction. In this paper we done the review of some load balancing algorithms based on different parameters. The main purpose of this paper is to help in design of some new algorithms in future by studying existing algorithms.

Keywords: Cloud computing, Load balancing, Performance metrics, and Existing techniques.

I. INTRODUCTION

From past few years cloud computing is one of major advances in the history of computing. It is the recent trend in the IT industry in which computing and data processing is done at large datacenters instead of desktops and portable PC's. Cloud is a term used for "INTERNET". The internet can be represented in a format as like a cloud. Cloud computing is simply an internet computing. It is made up by aggregating two terms in the field of technology. First term is a cloud and second term is computing. Cloud is a pool of heterogeneous resources. It is a mesh of huge infrastructure. Infrastructure refers to both the applications delivered to end users as services over the Internet and the hardware and system software. Computing is simply the utilization of internet to provide proper services to the people and organizations [1].

According to NIST(National Institute of Standards and Technology), "*cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources i.e. network servers, storage applications and services*" [2].

It is a new technology. It provides online resources and online storage to the users. It provides all the data which the user can access at lower cost but on paid basis. The aim of cloud computing is to achieve maximum resource utilization with higher availability at minimized cost. Cloud computing is a distributed computing which provides wide range of users with distributing access to hardware or software over the internet.

Services provided by cloud: - Three main types of cloud services [3] [4] are as follows.

1. **Software as a service (SaaS) [4]** - How software is delivered. It refers to software that is accessed via web browser and is paid. Example: Google apps.
2. **Platform as a service (PaaS)** – It provides development environment as a service. User deploys their apps on cloud and controls them. Example: Google app engine.
3. **Infrastructure as a service (Iaas)** – Also known as Hardware as a service. With this consumer get access to the infrastructure to deploy their stuff. But they do not manage and control the infrastructure1. Example: Compute cloud (EC2), Simple storage service.

Cloud can be deployed as public cloud, private cloud, community cloud and hybrid cloud. Cloud has some characteristics like: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. But there are some problems associated with cloud. Load balancing is one of them. The primary purpose of a cloud system is that its client can utilize the resources efficiently to gain some benefits. A resource allocation process is required to avoid overutilization or underutilization of resources which may affect services of cloud. Resource allocation and efficient scheduling improves the performance of system [4].

In this paper we present a review various load balancing algorithms in cloud environments. We provide an overview of these algorithms and discuss their properties.

II. LOAD BALANCING

As, we know that in today's world demand for the network resources growing day by day thus, creating a load which needs to be balanced. For maximum user satisfaction in the terms of high throughput and minimum response time with better utilization of resources "Load Balancing" is done. Load Balancing [5] is a one of the major challenge of cloud computing which needs to be focused. It is a process of reassigning the total load of system to an individual node in such

a way that no node should be overloaded or under loaded i.e. if the load on a node is more than a threshold value then this is to be shifted to a node with the fewer loads. In cloud it is used to distribute processing load evenly to all nodes. Load here can be network load, amount of memory used, CPU load etc.

Apart from this it also saves energy consumption thus helps in promoting towards energy efficient i.e. Green computing [6]. The main objective of load balancing in cloud computing is to improve performance in cloud, backup plan in case of system failure, reduce associated costs and response time that leads to user satisfaction. In cloud computing load balancing uses for balancing load on virtual machine and network resources. For this purpose a load balancer is used which receives task request from several users and distributes evenly it to datacenters [7]. Various load balancing algorithms are designed by various researchers which can be categorized as follows:

1) Depending on who initiate the process: [8]

a) Sender initiated: - If sender initiates the load balancing algorithm.
b) Receiver initiated: - If it is initiated by receiver.
c) Symmetric: - It is the combination of both sender initiated and receiver initiated.

2) On the basis of Current state of system:

a) **Static load balancing: -** It requires the prior knowledge of the system in terms of processing power, memory, and data as per user requirements. The decision of shifting load does not depend on current state of system. There are many drawbacks regarding this type is like sudden failure of system resource and tasks, also the task is assigned to processors or machines only after it is created and that task cannot be shifted to other machine during its execution for balancing the load. These algorithms are suitable for homogeneous environments. These algorithms are non preemptive so each machine has at least one task for execution. [8]

b) **Dynamic load balancing: -** It overcomes the drawbacks of static approach as decision of balancing the load depends on current state of system. Any prior knowledge of system is not required. It allows a process to move from an over utilized machine to underutilized for faster execution. This allows preemption which is not supported by static approach [9]. There algorithms are complex but they gives better performance and fault tolerance. These algorithms are more flexible than static algorithms and can easily adopt the change and provide better results so more suitable for heterogeneous and dynamic environment also. Dynamic load balancing algorithms can also be of distributed and non distributed nature.

In **distributed** not only a single node but all nodes in system takes part in load balancing by proper allocating the resources. In this each node has to interact with other nodes so, a lot of messages are produced. This distributed technique may be of cooperative in which all nodes works side by side to achieve a common goal and non- cooperative in which each node works independently towards the goal.

In **non-distributed** instead of all the nodes a particular node or a group of nodes is responsible for resource allocation or task scheduling. These algorithms are further classified into centralized and semi distributed. For centralized load balancing algorithm a single node called central node is responsible for balancing load in whole system and all other nodes have to interact with this node. In semi distributed clusters are formed. The nodes of system are partitioned into clusters where in each cluster the load balancing is of centralized form. A central node in each cluster is elected by using an appropriate election technique which takes care of balancing all nodes in cluster. Thus for a system central nodes of clusters are responsible for achieving the goal.

Some policies are used for dynamic load balancing algorithm [8]:

1) **Information policy:** What information is required and how this information is collected.
2) **Location policy:** Selection of destination node for transferring the task in load balancing algorithm.
3) **Resource type policy:** Defines types of resources which are available during load balancing.
4) **Selection policy:** Used to find out the task which transfers from overloaded node to a free node.
5) **Load estimation policy:** It estimates the total workload of a node in system.

III. WHY LOAD BALANCING IS DONE?

In cloud it is a mechanism which distributes the workload evenly across all the nodes to satisfy the user with better resource utilization. It also helps in ensuring that no node is overloaded which improves the performance of system. Apart from this proper load balancing helps in achieving green computing in following ways: [10]

1) Reducing energy consumption: - It helps in balancing workload against all nodes thus prevents them from overheating which Reduces the amount of energy used.

2) Reducing carbon emission: - Energy consumption and carbon emission are related to each other. More the energy consumed more the carbon footprint. As we know that load balancing helps in reducing energy consumption so Carbon emission also maintained.

Load balancing is done so that each machine in cloud does same amount of work therefore increasing the throughput. It is the one of the factor which determines the performance of a cloud.

IV. METRICS FOR LOAD BALANCING

There are some parameters which are considered important for a load balancing algorithm [8]:

1) Throughput: - Total no. of tasks that completed the execution. A high throughput is required to improve performance of system.

2) Overhead: - The amount of overhead that is produced by load balancing algorithm. This overhead is due to movement of tasks and inters processor communication. It should be minimized.

3) Response time: - The time taken by load balancing algorithm to respond for a particular task. This parameter should be minimized.

4) Resource utilization: - It is the degree to which resource is utilized. A good load balancing algorithm should provide maximum resource utilization.

5) Fault tolerant: - It is the ability of a system to perform in uniform manner even in case of system failure. The system implementing load balancing algorithms should be fault tolerant.

6) Migration time: - It is the time taken by one task to move from one system to other. It is an overhead which cannot be removed but should be minimized.

7) Scalability: - It is the ability of system to perform with finite no. of nodes in system.

8) Performance: - It is the effectiveness of system. For a system to be effective the response time of task be reduced while maintaining acceptable delay. If all above parameters are satisfied then performance also improved.

V. REVIEW OF LOAD BALANCING ALGORITHMS

Several load balancing have been proposed. Some of them are discussed here as follows.

- **Dynamic energy aware capacity provisioning:** - The algorithm proposed in paper [11] is based on saving energy in data centers by dynamically adjusting data center capacity by turning off unused machines or to set them to a power saving state. This algorithm uses a model predictive control (MPC) which minimizes the total energy cost while meeting performance objective in terms of task scheduling delay. Parameters used in this algorithm are Resource utilization, task scheduling delay (SLA cost), machine reconfiguration cost, and electricity price. It also predicts the future usage of resources in system i.e. CPU and memory. ARIMA model is used for prediction. Controller is responsible for reducing total operational cost of system. Which is the sum of SLA cost and energy cost. A high task scheduling delay affects the performance of some tasks. MPC algorithm adjusts the no. of servers/VM's to track the optimality condition while considering switching cost of machines. Bottleneck resource (a resource having high utilization) plays important role in this. Capacity provisioning module decides which machine to be added or removed based on certain criterion like usage of machine and its location. This algorithm gives the better results in saving energy with high performance and average resource utilization. However this algorithm considers all machines to be homogeneous with identical resource capacities.
- **Virtual machine placement algorithm:** - In [12], Virtual machine (VM) placement is the process of selecting the most suitable server in large cloud data centers to deploy newly-created VMs. Several approaches have been proposed to find a solution to this problem. Existing solutions only consider a limited number of resource types, thus resulting in unbalanced load or in the unnecessary conditions. Here, we propose an algorithm, called Max-BRU that improves the utilization of resources and maintains the usage of resources across multiple dimensions. This algorithm uses multiple resource-constraint metrics that help to find the most suitable server for deploying VMs in large cloud data centers. Parameters used in this algorithm are Virtual machines, resource utilization ratio (RU), resource balance ratio (RB). Resource may be CPU capacity, Memory and Network bandwidth. The main advantage of Max- BRU algorithm is that First it increases the resource utilization by minimizing the amount of physical servers used. Second, it effectively uses the multiple type of resources.
- **Round robin load balancing:** - In [13], the algorithm is proposed called Round robin, which uses the time slicing mechanism. As name implies this algorithm works in the round manner in which each node has given a time slice and has to wait for their turn. The time is divided into interval is allotted to each node in which nodes have to perform their task. It uses random selection procedure where first node is selected randomly and jobs are allocated to other nodes in a round robin fashion. The key point of this algorithm is that it yields no starvation and gives a faster response in case of equal workload distribution among processes. But as different processes have different processing times, therefore at any time some nodes may be heavily loaded while others remain idle and underutilized.
- **Central load balancing decision model:** - In [14] proposed algorithm is called as CLBDM. It is based on the human administrator point of view. It is a combination of the Round Robin Algorithm and session switching at the application layer. Round Robin is a used for the load balancing. In round robin algorithm, it assigns the task to the node with the least number of loads. In this algorithm a threshold time is considered. If connection time between the client and the node in the cloud is greater than the threshold time then there is an issue. If an issue is found, then the connection between client and node will be terminated and the task will be moved to another node using the regular Round Robin rules. The key point of this algorithm is that it is superior to Round Robin Algorithm as automated tasks forwarding reduces the need for a human administrator. But there are some drawbacks of this algorithm are that if this algorithm fails to work properly then whole process/system fails.
- **Min- Min load balancing: -** In the proposed algorithm known as Min- Min [15], initially there is a task set which is not assigned to any of node. For all the available nodes the minimum completion time is calculated. On finding the minimum time the task having the completion time minimum is chosen and assigned to the respective node. The execution time of all other tasks available in that machine is updated and the task gets discarded from the available task set. Once all the tasks have been assigned to proper machine this process is repeated. The advantages of this algorithm are that it is a simple and fast algorithm yields improved

performance. The algorithm works better in the situation where small tasks are more in number than larger tasks. However, the main drawback is it assigns the smaller task first as, smaller tasks will get executed first, while the larger tasks keeps on in the waiting stage ,which will finally results in poor machine use. It leads to starvation and it is less fault tolerant and Less Scalable.

- **Max- Min load balancing: -** Max-min load balancing algorithm [16] is similar to the min-min algorithm except the following: First it finds the minimum execution times, then the maximum value is selected which is the maximum time amongst all tasks on the resources. Then according to the maximum time, the task is scheduled on the corresponding machine. The execution time for all other tasks is updated on that machine and the assigned task is removed from the list of tasks that are to be assigned to the machines. This algorithm has advantages over Min-Min algorithm where smaller tasks are in high numbers as compared to that of larger ones. For e.g. if in a task set only a single larger task is presented then Max- Min algorithm runs smaller tasks concurrently along with larger task. This algorithm gives high throughput and optimum resource utilization. The algorithm suffers from starvation where the tasks having the maximum completion time be executed first while the tasks having the minimum completion time have left behind.
- **Load balancing Min- Min algorithm:** - In [17] an algorithm called Load Balancing Min-Min (LBMM) which is based on the Opportunistic Load Balancing algorithm (OLB). OLB has the aim to keep each node busy in cloud and is also static algorithm. The drawback of OLB is that is does not considers the execution time of a node. An improvement in OLB is done by LBMM which is a three layered architecture. At the first level the request manager receives the task and assigns it to one service manager which is at second level. When the service manager accepts the request, it divides it into subtasks. A service manager assign the subtask to a service node which executes the task by using different attributes such as the remaining CPU space, remaining memory and the transmission rate. The main drawback of this algorithm is that it is slower than other algorithms because work may pass through 3 layers before it has to be processed.
- **Active clustering:** - Active Clustering is a clustering based algorithm which introduces the concept of clustering in cloud computing. In cloud computing there are many load balancing algorithms available. The performance of an algorithm can be enhanced by making a cluster of nodes [18]. A group of cluster can be made. Active clustering group similar nodes together and then processes these groups. The process of creating a cluster revolves around the concept of match maker node. In this process, a node selects a neighbor node known as match maker node which is of a different type. This match maker node Searches for its neighbor node which is of same type as of initial node and makes connection with that node. Finally the matchmaker node gets detached. This process is repeated again and again. This increases the performance of the system with high availability of resources, throughput also increases. This increase in throughput is due to the efficient utilization of resources.

VI. CONCLUSION

Load balancing is the one of important issue faced using cloud computing. We have to be dealt with the problem of overloaded nodes. The load has to be distributed evenly throughout the nodes. Load balancing helps in optimal utilization of resources, increased performance with energy consumption. In this review we have discussed about various load balancing algorithms and the issues related with them which must be taken into account while designing new algorithms. Existing load balancing techniques mainly focus on reducing associated overhead, service response time and improving performance etc. some of the techniques has considered the energy consumption and carbon emission factors. Different performance mercies are defined like response time, migration time, scalability, throughput etc. Future work can be done by exploring new load balancing algorithms which balances the load much better and also helps in green computing.

REFERENCES

[1] B. P. Rima, E. Choi, and I. Lumb, "A Taxonomy and Survey of Cloud Computing Systems", Proceedings of 5th IEEE International Joint Conference on INC, IMS and IDC, Seoul, Korea, August 2009, pages 44-5l.

[2] Mell, Peter and Grance, Tim, "The NIST definition of cloud computing", National Institute of Standards and Technology, 2009, vol53, pages50, Mell2009

[3] Anthony T.Velte, Toby J.Velte, Robert Elsenpeter, Cloud Computing A Practical Approach, TATA McGRAW-HILL Edition 2010.

[4] Sahu, Yatendra and Pateriya, RK, "Cloud Computing Overview with Load Balancing Techniques", International Journal of Computer Applications, 2013, vol. 65, Sahu2013

[5] N. Sran and N. Kaur, "Comparative Analysis of Existing Load Balancing Techniques in Cloud Computing", *International Journal of Engineering Science Invention, 2(1)*, January 2013.

[6] Baliga, R. W. A. Ayre, K. Hinton, and R. S. Tucker, "Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport", Proceedings of the IEEE, Vol. 99, No. 1, January 2011, pages 149-167.

[7] S.S. Moharana, R.D. Ramesh and D. Powar, "Analysis of Load Balancers in Cloud Computing", *International Journal of Computer Science and Engineering, 2(2)*, May 2013.

[8] Ali M. Alakeel, A Guide to Dynamic Load Balancing in Distributed Computer Systems, IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.6, June 2010.

[9] K. Al Nuaimi, N. Mohamed, M. Al Nuaimi and J. Al-Jarrodi, "A Survey of Load Balancing in Cloud Computing: Challenges and Algorithms", *Network Cloud Computing and Applications*, 2012, 137-142.

[10] Nidhi Jain Kansal, Inderveer Chana, "Cloud Load Balancing Techniques: A Step Towards Green Computing", IJCSI, Vol. 9, Issue 1, January 2012.

[11] Qi Zhang , Shuo Zhang, Raouf Boutaba and Mohmed Faten Zhani, "Dynamic energy capacity provisioning for cloud computing environments" In Proceedings of the International Conference on Autonomic Computing (ICAC), 2012.

[12] Nguyen Trung Hieu, Mario Di Francesco, and Antti Yl¨a-Jaaski, "A Virtual Machine Placement Algorithm for Balanced Resource Utilization in Cloud Data Centers", *International Conference on Cloud Computing*,2014.

[13] Nusrat Pasha, Dr. Amit Agarwal and Dr.Ravi Rastogi,"Round Robin Approach for VM Load Balancing Algorithm in Cloud Computing Environment" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, May 2014.

[14] Radojevic, B. and M. Zagar, "Analysis of issues with load balancing algorithms in hosted (cloud) environments." In proc.34th International Convention on MIPRO, IEEE, 2011

[15] T. Kokilavani and Dr. D.I. George Amalarethinam, "Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing"International Journal of Computer Applications Volume 20– No.2, pp.0975-8887, April 2011.

[16] Xiaofang Li, Yingchi Mao, Xianjian Xiao, Yanbin Zhuang, An Improved max - min task scheduling algorithm for elastic cloud", IEEE proceedings of International Symposium on Computer, Consumer and Control, 2014, pp-340-343

[17] Wang S., Yan K., Liao W. and Wang S. (2010) 3*rd International Conference on Computer Science and Information Technology*, 108-113.

[18] Randles M., Lamb D. and Taleb-Bendiab A. (2010) 24*th Inter-national Conference on Advanced Information Networking and Applications Workshops*, 551-556.

