ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Review on Load Balancing Algorithms in Cloud Computing

Shasmita Panigrahi[1], Dhanraj Poojary[2], Dr Murlidhar Dhanawade[3]

PG Student, Dept. of MCA, NCRD's Sterling Institute of Management Studies, Navi Mumbai, India[1]

PG Student, Dept. of MCA, NCRD's Sterling Institute of Management Studies, Navi Mumbai, India[2]

Professor, Dept. of MCA, NCRD's Sterling Institute of Management Studies, Navi Mumbai, India[3]

ABSTRACT: Cloud computing can be considered as Computing powered by the Internet that provides shared processing resources and data storage options. Since more users are opting for cloud-based computing the load is also eventually increased on the cloud. Load balancing is a technique of distributing the load over the different nodes. Cloud Computing uses various Allocation, Scheduling, and Load balancing algorithms to migrate incoming tasks or processes to unutilized or underutilized Virtual Machines (VM) for effective utilization of resources and to provide maximum throughput. Load Balancing is essential to distribute the dynamic workload evenly across the nodes and has turned out to be most significant part in distributed environment. It consists of various difficulties such as security, virtualization, fault tolerance, etc. This paper portrays an overview of diverse types of load balancing algorithms or techniques in cloud computing and a detailed study of their advantages, disadvantages, performance metrics and challenges.

KEYWORDS: Cloud Computing, Load Balancing, Virtual Machine (VM), Algorithm

I. INTRODUCTION

Cloud Computing is a new technology of large scale distributed computing and which aims to have shared resources and processing powers. According to the definition of NIST (National Institute of Standards and Technology)[1] "Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. network, server, storage and applications, and services) that can be rapidly provisioned and released with minimal management efforts or service provider interaction."

Cloud computing is a technology which provides various computing services like a server, database, networking, software analytics and many such services over the internet on a Pay-per-use-on-demand model [6]. With the major advancements and success of cloud, it is attracting more users and enterprises to store and process their data on the cloud thereby reducing the infrastructure costs and at the same time achieving data integrity, scalability, reliability, high productivity and security at minimal costs.

There are three basic types of cloud deployment model available namely private cloud, public cloud, hybrid cloud and two variations namely Virtual Private Cloud and Community Cloud[5].The cloud computing services are broadly categorized into three categories that are Iaas (Infrastructure as a service), Paas (Platform as a service) and Saas(software as a service).Since Cloud comprises of a huge number of physical and virtual servers, effective management of this infrastructure is a matter of concern [7].Cloud computing would surely serve as a next generation computation and will revolutionize the way development and deployment are implemented in distributed environment.

II. LOAD BALANCING

Load Balancing is the method to divide the workload a single computing resource has to perform across one or more servers, network interfaces, hard drives or other computing resources. Load balancing technique could be implemented with the help of hardware, software or a combination of both[17]. A load can be CPU load, memory capacity, delay or network load. Load Balancing optimizes the resource utilization, maximizes throughput, maintains workload and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

improves response time [18].Load Balancer accepts the incoming requests and using the various assignment algorithms assigns the task or request to the appropriate computing resource based on task completion time and availability.

A. *Goals of Load Balancing:*

Goals of Load Balancing are as follows [19]:

- To improve the overall performance significantly.
- To Maintain the Stability of the system.
- To make the system robust and flexible for future modifications and implications.
- To build a fault tolerant system using backup mechanisms to secure the system from crashes or failures.

B. *Classification of Load Balancing Algorithm:*

There are different types of load balancing algorithms available which can be used depending upon the type of load and could be classified into two categories[19]. The following section will describe these two main categories of load balancing algorithms which are as follows:

Based on Process Initialization:
The Load balancing algorithm can be classified into three categories depending upon who initiated the process [19]:

1. Sender Initiated: If the load balancing algorithm is initiated by the sender then it is known as Sender Initiated.
2. Receiver Initiated: If the load balancing algorithm is initiated by the receiver then it is known as Receiver Initiated.
3. Symmetric: It is the combination of both sender initiated and receiver initiated.

Based on System State
The load balancing algorithm can be classified into two categories based on the current state of the system [19]:

1. Static**:** The Static Load balancing algorithm is not dependent on the current state of the system and also requires the prior knowledge of the system. Assignment of a task is done before the execution of program i.e. compile time.
2. Dynamic**:** The Dynamic load balancing algorithm is dependent on the current state of the system and does not require prior knowledge of the system. It is based on redistribution of the process among the processors during the execution time. It consists of four policies namely Transfer Policy, Selection Policy, Location Policy and Information Policy [28].

III. LOAD BALANCING ALGORITHMS

1. Round Robin Algorithm
The Round Robin load balancing algorithm is one of the simplest technique for distributing and allocating incoming client requests across a cluster of processing nodes. Round Robin Load Balancer forwards the incoming request to the appropriate server from the list and the first server selected for allocation is randomly selected from the list [20]. The load balancer assigns the incoming request to the list of servers on a rotating basis or in a circular order [2]. When it reaches the end of the list, the load balancer loops back and traverse the list of server again. The Main benefit of using Round-Robin algorithm is that it is extremely easy and simple to implement. But the major drawback of this algorithm is it does not ensure accurate or efficient distribution of traffic since it assumes that all server are up and running, handling the same load and with same storage and computing capacities [20].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

1.1 Weighted Round Robin Algorithm
In this algorithm, a weight is assigned to each server based on certain criteria as specified by the administrator [20]. The proportion of weight assigned to the server acts as the determining criteria for the number of requests the server receives.

1.2 Dynamic Round Robin Algorithm
In this algorithm unlike in the weighted round robin algorithm the weight is assigned dynamically to the server based on real-time data gathered from the server's current load and idle capacity details [20].

2. Throttled Load Balancing Algorithm (TLB)
The Throttled Load Balancing Algorithm uses an index table containing virtual machines and their current status (busy or available). The requesting Client or server first initiates a request to the data centre to search for a suitable virtual machine (VM) to perform the task. The throttled load balancer then performs a scan of index table starting from the top until a VM is found with current state available or the entire index table is scanned [2]. If a VM is found, the id of VM is sent to the data centre or else the load balancer returns -1 to the data centre [21], [22]. Further, the load balancer continuously acknowledges the data centre of the new allocation in regular intervals.

2.1 Modified Throttled Load Balancing Algorithm
The modified throttled load balancing also maintains a list of index table of virtual machines (VM) like the Throttled algorithm. Also, the first VM is selected in the same manner. But during the next request arrives, the VM at index table next to already assigned VM is chosen based on the current state or availability of the VM. The main benefit of using modified variant of throttled algorithm is it gives better response time as compared to throttled algorithm. But it is observed that the state of some VM may change based on the allocation and deallocation of tasks. So it is not always beneficial to start searching for the next to already assigned VM [2].

3. Load Balance Min-Min Algorithm (LBMM)
The Load Balance Min-Min algorithm uses Min-Min Scheduling algorithm as its base which consists of a set of tasks and then resource which has minimum completion time for all tasks is selected, then the task with minimum size is selected and assigned to the appropriate resource. It consists of a three-level hierarchical framework with Request Manager at first level of architecture which is responsible for receiving the task and assigning it to the appropriate service manager in the second level. After receiving the request, service manager divides the tasks into sub-tasks at the second level of architecture and assigns the subtasks to appropriate service node which resides at the third level of execution of the task. This algorithm greatly improves the load unbalance of Min-Min and minimizes the overall execution time, but fails to specify the selection of node for a complicated task comprising of large scale computation [23].

3.1 Load Balance Improved Min-Min Scheduling Algorithm (LBIMM)
This algorithm also starts with the execution of Min-Min algorithm at the initial step. In next step, it selects the smallest task from the heavily loaded resource and performs calculations for the completion time of that particular task on all other resources. If the value is less than the value of makespan then the task is reassigned to the resource that generated it and eventually updates the ready time of both resources [2].

3.2 User-Priority Aware Load Balance Improved Min-Min Scheduling Algorithm (PA-LBIMM)
This Algorithm is incorporated with LBIMM to produce PA-LBIMM. This algorithm initially divides all the tasks into two groups G1 and G2 [2]. Group G1 consists of high priority user's task and group G2 consists of ordinary or low priority user's task. The higher priority task of group G1 is assigned first using the Min-Min algorithm to the high priority qualified resource set. At the end load of all the resources are optimized to create a final schedule. The algorithm is more focused on makespan, load balancing and user priority with a drawback of not able to consider the deadline for each task [24].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

4. Two Phase Load Balancing Algorithm

This algorithm is a combination of OLB (Osmosis Load Balancing) which works on the principle of osmosis to reschedule the incoming tasks in the virtual machines [25] and LBMM. In the initial phase the OLB Scheduling manager assigns the task to the service manager and in the second phase, LBMM choose the suitable service node to execute the subtask by the service manager with the only drawback that it could be used in the static environment.

5. Artificial Bee Colony Algorithm (ABC)

This algorithm is proposed by J. Yao and J. H. He[26] based on the requirements and characteristics of cloud computing environments. In this type of strategy, hundreds of thousands of simultaneous requests of the same type are queued on the same server. Consequently, it raised local resource intensive phenomenon and deteriorated load balancing.

5.1 Improved Artificial Bee Colony Algorithm (IABC)

This algorithm was introduced to overcome the failure or shortcomings of ABC Algorithm. This algorithm replaced another type of requests with the next server request which eventually changes the type of the request thereby ending the accumulation of request and improves the throughput of the system.

6. Ant Colony Optimization

This algorithm consists of a model which considers the behavioural aspects of ant. Ants have a very limited memory and exhibit to have a larger component. Acting as a collective unit ants manage to perform a variety of incoming tasks which are complex with great reliability, efficiency, and consistency [27].The ants in the proposed algorithm continuously originate from the head node and scan the entire network making bidirectional movements to find the under loaded and overloaded nodes. ACO consists of two types which are Foraging Pherome and Trailing Pherome. To limit the ants, they commit suicide once they find target node.

IV. METRICS

Following are the metrics used in load balancing techniques:

- Scalability- It is the ability of an algorithm to perform the load balancing efficiently with any given number of nodes.
- ResourceUtilization- It is used to check the utilization of resources.
- FaultTolerance- It is the time required to migrate jobs from one node to other.
- ResponseTime- It is the metric used to determine the amount of time taken to respond by a particular load balancing algorithm.
- Throughput- It is used to calculate the number of tasks which have completed execution.
- Overhead- It is the metric used to determine the amount of overhead implemented in the load balancing algorithm which comprises of overhead due to movement of task and inter process communication.
- Performance- It is used to check the efficiency of the system.

V. COMPARISON AND ANALYSIS

Load Balancing Methodology	Parameter	Merits	Demerits	Variations
Round Robin Algorithm	Allocation Based on a Circular Order	Easy and simple to implement	Does not ensure accurate or efficient allocation	-Weighted Round Robin -Dynamic Round Robin
Throttled Load Balancing Algorithm	Allocation based on state found in VM	Efficient Allocation to under loaded VMs	Allocation may dither due to change	-Modified Throttled Load

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

	index table		in state of VMs	Balancing Algorithm
Load Balance Min-Min Algorithm	Based on two step(min-min) decisive factors for allocation	Improves the load allocation and overall execution time	Inefficient to allocate nodes for complicated task	-LBIMM -PA-LBIMM
Two Phase Load Balancing Algorithm	Based on two phase allocation(OLB and LBMM)	Improved Execution and efficiency	Can be used only in static environment	None
Artificial Bee Colony Algorithm	Accommodate hundreds of request in queue of same server	Local resource intensive phenomenon	Deteriorated load balancing	-Improved Artificial Bee Colony Algorithm
Ant Colony Optimization	Allocation based on behavioural aspect of ants	Improved Allocation execution and efficiency	Cannot provide precise path for allocation	Max Min Ant System

VI. CHALLENGES

Although Cloud computing is widely used and adopted across globe but still its research is in its early stages and some challenges remain unsolved in particular the load balancing challenges.

- Automated Service Provisioning: One of the key features of cloud computing is elasticity which is acquiring and releasing of resources [3]. Service provider has to effectively allocate and deallocate resources while reducing the operational costs. There is Proactive and reactive resource control [4].
- Management of Energy: Saving energy is a key point that allows a global economy to implement set of resources supported by reduced providers rather than each one having its own resource [3].It has been estimated that the cost of powering and cooling accounts for 53% of total expenditure [29].
- Data Management: Storage of data across the network has grown exponentially and management of data has become a major challenge, thus raising the question of how data can be optimally stored.
- Migration of Virtual Machines: Through the concept of virtualization an entire machine could be seen as an individual file or set of files. To unload a heavily loaded machine is it possible to move a virtual machine between physical machines [3]. The main agenda is to distribute and balance the overall load in datacenter.

VII. CONCLUSION

Cloud Computing has been widely accepted and adopted by many industries and users across the globe though there are many existing issues and challenges discussed in the paper. Central to this issue lies the issue of load balancing which is required to distribute the workload evenly between all the nodes to achieve proper resource utilization and produce a good throughput of the system, attain a high user satisfaction and reduce waiting time. Major emphasis is given to the study of load balancing algorithm and their variants, working, merits, demerits, comparison followed by challenges in load balancing.

REFERENCES

1. Peter Mell, Timothy Grance, "The NIST Definition of Cloud Computing", Special Publication 800-145.
2. Geetha Megharaj, Dr. Mohan K.G., "A Survey on Load Balancing Techniques in Cloud Computing", IOSR Journal of Computer Engineering, Volume 18, PP 55-61.
3. Deepak B S, Shashikala S V, Radhika K R, "Load Balancing Techniques in Cloud Computing: A Study", International Journal of Computer Applications (0975 – 8887).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

4. Qi Zhang, Lu Cheng, Raouf Boutaba, "Cloud computing: state-of-the-art and research challenges", Journal of Internet Services and Applications, Volume 1, PP 7–18.
5. Types of Cloud, http://www.service-architecture.com/articles/cloud-computing/types_of_clouds_in_cloud_computing.html
6. Jeep Ruiter, MartijN Warnier, "Privacy Regulations for Cloud Computing Compliance and Implementation in Theory and Practice".
7. Dancho Danchev,"Building and implementing a successful Information Security Policy", http://www.windowsecurity.com/pages/security-policy.pdf
8. Akanksha Tripathi, Prof. Sandeep Raghuwanshi, "Cloud Task Scheduling in Hybrid Environment using Ant Colony Optimization", International Journal of Advanced Engineering and Global Technology I, Volume 03, Issue-07, July 2015.
9. Amazon Elastic Computing Cloud, aws.amazon.com/ec2
10. Cloud Computing and Hybrid Infrastructure from GoGrid, http://www.gogrid.com
11. FlexiScale Cloud Comp and Hosting, www.flexiscale.com
12. Google App Engine, http://code.google.com/appengine
13. Windows Azure, www.microsoft.com/azure
14. Salesforce CRM, http://www.salesforce.com/platform
15. Dedicated Server, Managed Hosting, Web Hosting by Rackspace Hosting, http://www.rackspace.com
16. SAPBusinessByDesign,http://www.sap.com/sme/solutions/businessmanagement/businessbydesign/index.epx
17. Aswathi Vandana P., Nandhini A., Saravana Balaji B., Dr.N.K.Karthikeya, "An Overview on Performance Issues in Cloud Computing", International Journal of Engineering Research & Technology, Vol. 2, Issue 9, September 2013.
18. R. Shinmonski. "Windows 2000 & Windows Server 2003.Clustering and Load balancing", Emeryville, McGraw-Hill professional publishing, CA, USA, P2, 2003.
19. David Escalante and Andrew J. Korty, "Cloud Services: Policy and Assessment", EDUCAUSE Review, Vol. 46, July/August 2011.
20. Round Robin load balancing technique, https://www.nginx.com/resources/glossary/round-robin-load-balancing/
21. B. Wickremasinghe, "CloudAnalyst: A CloudSim-based tool for modelling and analysis of large scale cloud computing environments", MEDC project report, 22(6), 433-659, 2009.
22. B. Wickremasinghe, R. N. Calheiros, and R. Buyya, "Cloud analyst: A CloudSim-based Visual Modeller for Analysing Cloud Computing Environments and Applications", Proc. 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), 446-452, 2010.
23. Shu-Ching Wang, Kuo-Qin Yan, Wen-Pin Liao, and Shun-Sheng Wang, "Towards a Load Balancing in a Three-level Cloud Computing Network", Proc. 3rd International Conference on Computer Science and Information Technology (ICCSIT), 108-113, 2010.
24. H. Chen, F. Wang, N. Helian, and G. Akanmu, "User-priority guided Min-Min scheduling algorithm for load balancing in cloud computing", Proc. National Conference on Parallel Computing Technologies (PARCOMPTECH), 1-8, 2013.
25. B. Mallikarjuna, P. Venkata Krishna, "OLB: A Nature Inspired Approach for Load Balancing in Cloud", The Journal of Institute of Information and Communication Technologies of Bulgarian Academy of Sciences, Volume 15, Issue 4, Nov 2015.
26. J. Yao, and J. H. He, "Load balancing strategy of cloud computing based on artificial bee algorithm", Proc. 8th International Conference on Computing Technology and Information Management (ICCM), 185-189, 2012.
27. Ratan Mishra, Anant Jaiswal, "Ant colony optimization: A Solution of load balancing in cloud", International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.2, April 2012.
28. Ali M. Alakeel, "A Guide to Dynamic Load Balancing in Distributed Computer Systems", International Journal of Computer Science and Network Security(IJCSNS), VOL.10, No.6, June 2010.
29. J.Hamilton, "Cooperative Expendable Micro-Slice Servers (CEMS): Low Cost, Low Power Servers for Internet-Scale Services", Proc. of CIDR.

BIOGRAPHY

Miss. Shasmita Panigrahi is a Post Graduate student of Master of Computer Application (MCA), College of NCRD's Sterling Institute of Management Studies, Mumbai University.

Mr. Dhanraj Poojary is a Post Graduate student of Master of Computer Application (MCA), College of NCRD's Sterling Institute of Management Studies, Mumbai University.

Dr. Murlidhar Dhanawade is a professor in Master of computer Application Department, College of NCRD's Sterling Institute of Management Studies, Mumbai University.