A Study of Various Load Balancing Techniques in Cloud Computing and their Challenges

Vinod K. Lalbeg,
Asst. Prof.
Neville Wadia Institute Management Studies &Research, Pune-1
vklalbeg@yahoo.com
Co-Author: Dr. Minesh Ade (Ph.D. Guide)

ABSTRACT

Cloud Computing is an emerging distributed computing paradigm. A framework which enables convenient, on-demand access to network, aiming to share data, calculations and service transparently over a shared pool of scalable network of nodes. Load balancing is one of the main challenges in cloud computing which is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overloaded. It helps in optimal utilization of resources and hence in enhancing the performance of the system. There are few existing scheduling algorithms that can maintain load balancing and provide better approaches through efficient job scheduling and resource allocation techniques as well. It becomes necessary to utilize the available resources efficiently in order to maximum profits with optimized load balancing algorithms. This paper discusses some of the load balancing algorithms in cloud computing and the various challenges faced by the organizations using them.

Keywords: Distributed computing, load balancing, algorithms, scheduling, dynamic

1. Introduction:

A Cloud computing is the paradigm of large scale distributed computing.Cloud computing provides the scalable IT resources such as applications and services, as well as the infrastructure on which they operate, over the Internet, on pay-per-use basis to adjust the capacity quickly and easily. Cloud Services allows individuals and businesses to use software and hardware that are managed by third party at remote locations. It helps to accommodate changes in demand and helps any organization in avoiding the capital costs of software and hardware [2] [3]. Thus, Cloud Computing is a framework for enabling a suitable, on-demand network access to a shared pool of computing resources (e.g. networks, servers, storage, applications, and services).These resources can be provisioned and de-provisioned quickly with minimal management effort or service providerinteraction. This further helps in promoting high availability[4]. Due to the exponential growth of cloud computing, it has been widely adopted by the industry and there is a rapid expansion in data-centers. The distributed computers provide on-demand services. Services may be of software resources (e.g. Software as a Service, SaaS) or physical resources (e.g. Platform as a Service, PaaS) or hardware/infrastructure (e.g. Hardware as a Service, HaaS or Infrastructure as a Service, IaaS). Amazon EC2 (Amazon Elastic Compute Cloud) is an example of cloud computing services [5].

The NIST defines cloud computing as:

"A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [6].

Cloud computing is still in evolving stage faces various problems. Some of the problems are

- Ensuring proper access control (authentication, authorization, and auditing)
- Network level migration, so that it requires minimum cost and time to move a job
- To provide proper security to the data in transit and to the data at rest.
- Data availability issues in cloud

- Data lineage, data provenance and inadvertent disclosure of sensitive information is possible
- While the most prevalent problem in Cloud computing is the problem of load balancing.

1.1 Cloud Computing Architecture

There are basically three layers that constitutes to the Cloud Computing Architecture. It provides three basic services viz. SaaS, PaaS and IaaS.

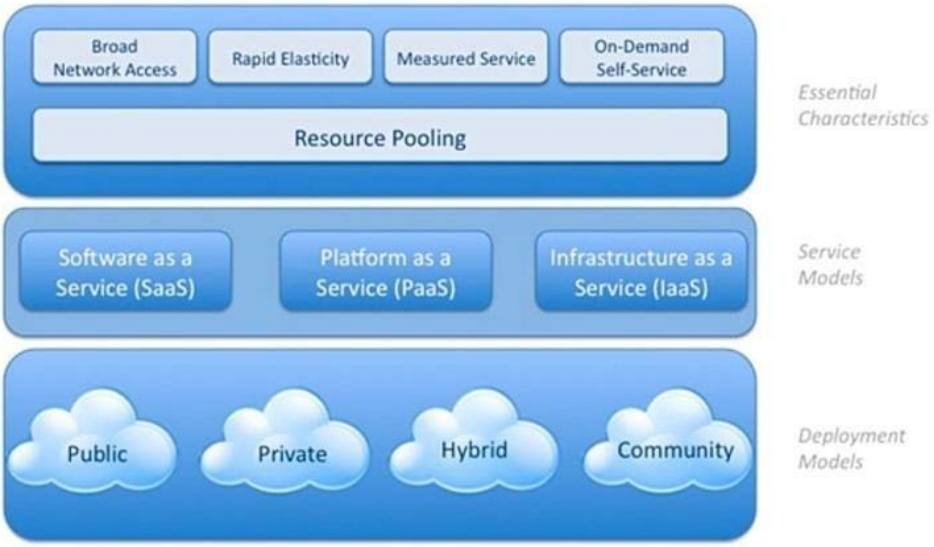

Figure 1 : Cloud Computing Architecture

2. Virtualization and Binding

Resource virtualization is at the heart of most cloud architectures. The concept of virtualization allows an abstract, logical view on the physical resources and includes servers, data stores, networks and software. The basic idea is to pool physical resources and manage them as a whole. Individual requests can then be served as required from these resource pools. For example, it is possible to dynamically generate a certain platform for a specific application at the very moment when it is needed. Instead of a real machine, a *virtual machine* is used [7]. There are two types of virtualizations:

a) Full Virtualization

In full virtualization the entire installation of one system is done on other system. Due to this all the software that is present in actual server will also be available in virtual system. This also helps in sharing of computer system among multiple users and emulating hardware located on different systems.

b) Para Virtualization

In Para Virtualization multiple operating systems are allowed to run on a single system by using system resources like memory and the processor. VMware software helps in achieving Para Virtualization. Here complete services are not fully available, but partial services are provided. Para virtualization helps in Disaster recovery, migration and capacity management.

3. Load Balancing

Load Balancing is a method to distribute workload across one or more servers, network interfaces, hard drives or other computing resources. Typical datacenter implementations rely on large, powerful computing hardware and network infrastructure, which are subject to the usual risks associated with any physical device, including hardware failure, power and/or network interruptions, and resource limitations in times of high demand. Load balancing in the cloud

differs from classical thinking on load-balancing architecture and implementation by using commodity servers to perform the load balancing.

Load balancing is used to make sure that none of your existing resources are idle while others are being utilized. To balance load distribution, you can migrate the load from the source nodes (which have surplus workload) to the comparatively lightly loaded destination nodes. This provides for new opportunities and economies-of-scale, as well as presenting its own unique set of challenges. (Mishra, et al., 2012) [8]

There are mainly two categories of load balancing algorithms:

3.1 Static

In static algorithm the network traffic is divided evenly among the servers or nodes. Static algorithm requires having prior knowledge of system resources, so that the decision of shifting of the load does not depend on the current state of the system. Static algorithm suits system which has low variation in load.

3.2 Dynamic

In dynamic algorithm the server which is either idle or having least load in the whole network or system is searched and preferred for assigning load. Here current state of the system is used to make decisions to manage the load. But for this real time communication with network is needed which can increase the traffic in the system.

4. Load Balancing Algorithms

We classified load balancing algorithm in two main types that are Static load balancing and Dynamic load balancing.

Round Robin (RR): In 2009, B Sotomayor et al [9] introduced a static well-known load balancing technique called Round Robin, in which all processes are divided amid all available, processors. The allocation order of processes is maintained locally which is independent of the allocation from the remote processor. In this technique, the request is sent to the node having least number of connections, and because of this at some point of time, some node may be heavily loaded and other remain idle [9]. This problem was solved by Central Load Balancing Decision Model (CLBDM)

Equally Spread Current Execution Algorithm: Equally spread current execution algorithm [10] process handle withpriorities. It distributes the load randomly by checking the size and transfers the load to that virtual machine which is lightlyloaded or handles that task easily and takes less processing, and gives maximize throughput. It is spread spectrum technique inwhich the load balancer spread the load of the job in-hand into multiple virtual machines.

Throttled Load Balancing Algorithm: Throttled algorithm [10] is completely based on virtual machine. In this the load balancer searches the right virtual machine which can execute the load easily and perform the operationswhich is given by the client or user. In this algorithm the client first requests the load balancer to find a suitable VirtualMachine to perform the required operation.

Biased Random Sampling:M. Randles et al. [12] investigated a distributed and scalable load balancing approach that uses random sampling of the system domain to achieve self-organization thus balancing the load across all nodes of the system. Here a virtual graph is constructed, with the connectivity of each node (a server is treated as a node) representing the load on the server. Each server is symbolized as a node in the graph, with each in degree directed to the free resources of the server. The load balancing scheme used here is fully decentralized, thus making it apt for large network systems like that in a cloud. The performance is degraded with an increase in population diversity.

Load Balancing Min-Min: In 2010, S C. Wang et al. [13] presented a dynamic load balancing algorithm called Load Balancing Min-Min (LBMM) technique which is based on three level frameworks. It begins with a set of all unassigned tasks. First, minimum completion time for all tasks isfound. Then among these minimum times the minimum value is selected which is the minimum time among all the taskson any resources. Then according to that minimum time, the task is scheduled on the corresponding machine. Then theexecution time for all other tasks is updated on that machine by adding the execution time of the assigned task to theexecution times of other tasks on that machine and assigned task is removed from the list of the tasks that are to beassigned to the machines. Then again the same procedure is followed until all the tasks are assigned on the resources.But this approach has a major drawback that it can lead to starvation [14].

Map Reduced Based Entity Resolution Load Balancing: In 2011, L. Colb et al [15] introduced the Map Reduced based Entity Resolution load balancing technique which is based on large datasets. In this technique, two main tasks are done: Map task and Reduce task which the author has described. For mapping task, the PART method is executed where the request entity is partitioned into parts. And then COMP method is used to compare the parts and finally similar entities are grouped by GROUP method and by using Reduce task. Map task reads the entities in parallel and process them, so that overloading of the task is reduced. In 2011, J Hu et al. [16] introduced a static scheduling strategy of load balancing on virtual machine resource. This technique considers the historical data and also the current state of system. Here, central scheduler and resource monitor is used. The scheduling controller checks the availability of resources to perform a task and assigns the same. Resource availability details are collected by resource monitor.

Ant Colony Optimization: In 2012, K. Nishant et al [17] introduced a static load balancing technique called Ant Colony Optimization. In this technique, an ant starts the movement as the request is initiated. This technique uses the Ants behavior to collect information of cloud node to assign task to the particular node. In this technique, once the request is initiated, the ant and the pheromone starts the forward movement in the pathway from the “head” node. The ant moves in forward direction from an overloaded node looking for next node to check whether it is an overloaded node or not. Now if ant find under loaded node still it move in forward direction in the path. And if it finds the overloaded node then it starts the backward movement to the last under loaded node it found previously. If ant finds the target node, it will commit suicide so that unnecessary backward movement is prevented.

Index Name Server Dynamic Load Balancing: In 2012, T. Yu Wu et al. [18] introduced a dynamic load balancing technique called Index Name Server to minimize the data duplication and redundancy in system. This technique works on integration of de duplication and access point optimization. To calculate optimum selection point some parameter are defined: hash code of data block to be downloaded, position of server having target block of data, transition quality and maximum bandwidth. Another calculation parameter to find weather connection can handle additional node or is at busy level B(a), B(b) or B(c). B(a) denote connection is very busy to handle new connection , B(b) denotes connection is not busy and B(c) denotes connection is limited and additional study needed to know more about connection.

Stochastic Hill Climbing Load Balancing: In 2012, B. Mondal et al [19] have proposed a load balancing technique called Stochastic Hill Climbing based on soft computing for solving the optimization problem. This technique solves the problem with high probability. It is a simple loop moving in direction of increasing value which is uphill. And this make minor change in to original assignment according to some criteria designed. It contains two main criteria one is

candidate generator to set possible successor and the other is evaluation criteria which ranks each valid solution. This leads to improved solution.

Honey Bee Behavior - Load Balancing [HBB-LB]: In 2013, D. Babu et al [20] proposed a Honey Bee Behavior inspired Load Balancing [HBB-LB] technique which helps to achieve even load balancing across virtual machine to maximize throughput. It considers the priority of task waiting in queue for execution in virtual machines. After that work load on VM calculated decides weather the system is overloaded, under loaded or balanced. And based on this VMs are grouped. New according to load on VM the task is scheduled on VMs. Task which is removed earlier. To find the correct low loaded VM for current task, tasks which are removed earlier from over loaded VM are helpful. Forager bee is used as a Scout bee in the next steps.

5. Challenges for Load Balancing Algorithms:

There are some qualitative metrics that can be improved for better load balancing in cloud computing [21][22].

Throughput: It is the total number of tasks that have completed execution for a given scale of time. It is required to have high through put for better performance of the system.

Associated Overhead: It describes the amount of overhead during the implementation of the load balancing algorithm. It is a composition of movement of tasks, inter process communication and inter processor. For load balancing technique to work properly, minimum overhead should be there.

Fault tolerant: We can define it as the ability to perform load balancing by the appropriate algorithm without arbitrary link or node failure. Every load balancing algorithm should have good fault tolerance approach.

Migration time: It is the amount of time for a process to be transferred from one system node to another node for execution. For better performance of the system this time should be always less.

Response time: In Distributed system, it is the time taken by a particular load balancing technique to respond. This time should be minimized for better performance.

Resource Utilization: It is the parameter which gives the information within which extant the resource is utilized. For efficient load balancing in system, optimum resource should be utilized.

Scalability: It is the ability of load balancing algorithm for a system with any finite number of processor and machines. This parameter can be improved for better system performance.

Performance: It is the overall efficiency of the system. If all the parameters are improved then the overall system performance can be improved.

6. Conclusion and Future Research

In this paper, the researcher has surveyed various load balancing techniques for cloud computing. I would also like to highlight the changes happening in the field of load balancing algorithms across the world. Distributing load dynamically among the servers and nodes to utilize maximum available resources and improving the performance of the system is the main objective of all the load balancing algorithms.

7. References :

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica and M. Zaharia, **"Above the Clouds: A Berkeley View of Cloud Computing"**, EECS Department, University of California, Berkeley, Technical Report No. UCB/EECS-2009-28, February 2009.

[2] R. W. Lucky, **"Cloud computing"**, IEEE Journal of Spectrum, Vol. 46, No. 5, May 2009

[3] M. D. Dikaiakos, G. Pallis, D. Katsa, P. Mehra and A. Vakali, **"Cloud Computing: Distributed Internet Computing for IT and Scientific Research"**, IEEE Journal of Internet Computing, Vol. 13, No. 5, September/October 2009

[4] G. Pallis, **"Cloud Computing: The New Frontier of Internet Computing"**, IEEE Journal of Internet Computing, Vol. 14, No. 5, September/October 2010

[5] Martin Randles, David Lamb, A. Taleb-Bendiab, A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing, 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops.

[6] Mell, P. a. (2009). Draft NIST Working Definition of Cloud Computing.

[7] Baun, C. (2011). Cloud Computing: Web-Based Dynamic IT Services. Springer Sotomayor, B., RS. Montero, IM.Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," in IEEE Internet Computing, Vol. 13, No. 5, pp.: 14-22, 2009.

[8] Nitika, Shaveta andGaurav Raj; "Comparative Analysis of Load Balancing Algorithms in CloudComputing", International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.

[9] P.Warstein, H.Situ and Z.Huang(2010), "Load balancing in a cluster computer" In proceeding of the seventhInternational Conference on Parallel and Distributed Computing, Applications and Technologies, IEEE.

[10] T.R.V. Anandharajan, Dr. M.A. Bhagyaveni,"Co-operative Scheduled Energy Aware Load-Balancing technique for anEfficient Computational Cloud", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.

[11] Wang, S-C., K-Q.Yan, W-P.Liao and S-S. Wang, "Towards a load balancing in a three-level cloud computing network", in the 3rd International Conference on Computer Science and Information Technology (ICCSIT), IEEE Vol. 1, pp. 108-113, July 2010.

[12] T. Kokilavani J.J. College of Engineering & Technology and Research Scholar, Bharathiar University, Tamil Nadu, India,"Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing", International Journal of Computer Applications (0975 – 8887) Volume 20– No.2, April 2011

[13] Kolb, L., A. Thor, and E. Rahm, E, "Load Balancing for MapReduce based Entity Resolution," in the 28th International Conference on Data Engineering (ICDE), IEEE, pp.: 618-629, 2012.

[14] J. Hu, J. Gu, G. Sun, and T. Zhao, "A Scheduling Strategy on Load Balancing of Virtual Machine Resources in Cloud computing Environment", Third International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), 2010.

[15] Nishant, K. P. Sharma, V. Krishna, C. Gupta, KP. Singh N. Nitin and R. Rastogi, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization", in 14th International Conference on Computer Modeling and Simulation (UKSim), IEEE, pp.: 3-8, March 2012. [18]. T-Y., W-T. Lee, Y-S.Lin, Y-S.Lin, H-L.Chan and J-S. Huang, "Dynamic load balancing mechanism based on cloud storage" in proc. Computing, Communications and Applications Conference (ComComAp), IEEE, pp: 102-106, January 2012.

[16] Brototi M, K. Dasgupta, P. Dutta, "Load Balancing in Cloud Computing using Stochastic Hill Climbing-A Soft Computing Approach", in proc. 2nd International Conference on Computer, Communication, Control and Information Technology(C3IT)-2012.

[17] Dhinesh B. L.D , P. V. Krishna, "Honey bee behavior inspired load balancing of tasks in cloud computing environments", in proc. Applied Soft Computing, volume 13, Issue 5, May 2013, Pages 2292-2303.

[18] Tushar Desai, JigneshPrajapati, "A Survey Of Various Load Balancing Techniques And Challenges In Cloud Computing" International Journal of Scientific & Technology Research Vol. 2, Issue 11, Nov. 2013, ISSN 2277-8616, pp.:158-161

[19] Foster, I., Y. Zhao, I. Raicu and S. Lu, "Cloud Computing and Grid Computing 360-degree compared," in proc. Grid Computing Environments Workshop, pp.: 99-106, 2008.

******