INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue2)
Available online at www.ijariit.com

Taxonomy of Load Balancing

Divya Garg
M.Tech. Student
CSE Department (CEC, Landran, Punjab)
divyagarg1900@gmail.com

Er. Anurag Jain
Associate Professor
CSE Department (CEC, Landran, Punjab)
er.anuragjain14@gmail.com

Abstract— Cloud computing is one of an emerging technology in the computer world these days. Virtualization is the concept behind cloud computing that allows sharing of resources like software, infrastructure and platform. Different people share different views regarding cloud computing. Some enterprises take a lot of advantages from cloud computing while other considers it being an unsafe place for storing critical information. Load balancing is one of an important issue in cloud computing. This paper is for those who want to understand the basics of load balancing.

Keywords— Cloud Computing, Load Balancing Issues, Load Balancing Approaches.

I. INTRODUCTION

Cloud computing is one of the most emerging technology in the world of computing these days. Cloud computing provides an environment to the users where data can be stored, retrieved and accessed very easily and with minimum human–service interaction.

Now a days, computing are becoming increasingly essential in every field. A lot of data is stored in a computer and the mass of data is increasing continuously. This increased data requires more storage devices to cope up with the ever increasing needs of organizations [11]

Cloud computing provides its users three service models namely: Software as a service (SaaS), Infrastructure as a Service (IaaS), and Platform as a Service (PaaS). Load balancing is an important concept in cloud computing. In the process of Load balancing [1] the total load to the individual nodes of the collective system is reassigned for making effective resource utilization and for improving the response time of the job, also concurrently a condition is removed in which some of the nodes are over loaded while many of them are under loaded. A load balancing protocol is somehow dynamic in nature and it doesn't consider the previous state or behaviour of the system, that is, it depends on the current behaviour of the system.

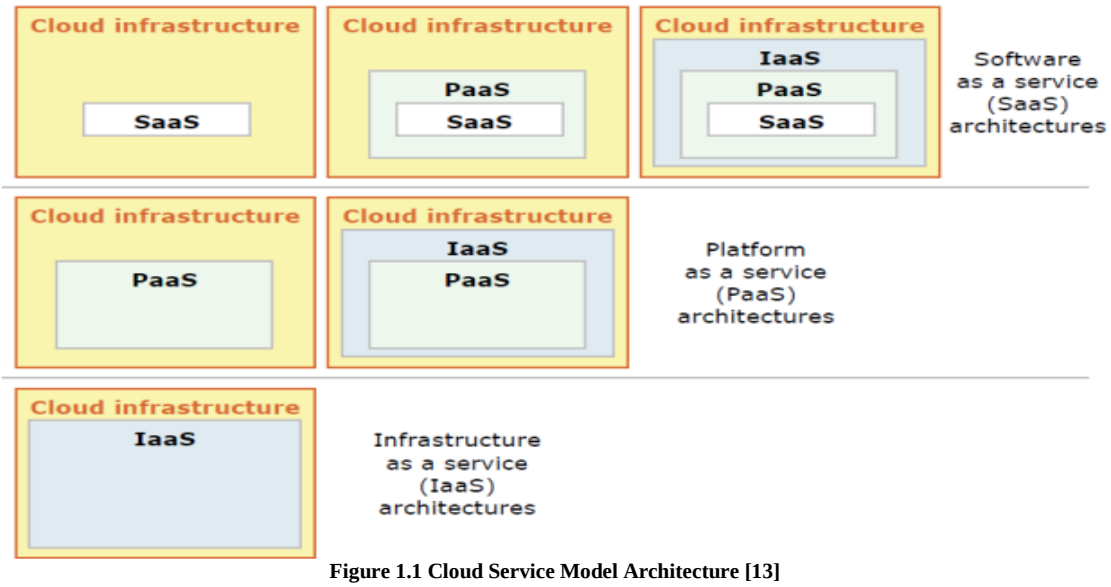

Figure 1.1 Cloud Service Model Architecture [13]

In this taxonomy we have tried to explain the basics of cloud computing. This paper is organized as section I, II, III, IV, V. In section II we have taken an overview of load balancing algorithm. In section III we have discussed the need of load balancing of machines In section IV different load balancing strategies and in section V we have discussed the challenges of load balancing in cloud. Lastly in section VII we discussed some existing load balancing algorithms.

II. OVERVIEW OF LOAD BALANCING IN CLOUD

Load balancing in clouds is a mechanism by which the excess dynamic local workload is evenly distributed across all the nodes. It helps in achieving the high performance of the system by ensuring user satisfaction, proper resource utilization, less burden on a single node of any system. Load balancing helps in utilizing the available resources optimally by minimizing the consumption of available resources [2].

Metrics for Load Balancing:

1. Throughput: Throughput is the maximum production of a system. A high throughput is required for the better performance of the system [10].

2. Fault Tolerance: Fault tolerance is the recovery from failure. Every load balancing algorithm must be fault tolerant so that in case of failure it can recover [10].

3. Migration time: Migration time is the time to migrate the jobs and resources from one node to another. It must be minimized so that the performance of the system can be enhanced [10].

4. Response Time: Response time is the amount of time an algorithm takes to response to a task in a system. It should also be minimized for better performance of a system [10].

5. Scalability: Scalability is the ability of a system to increase the no of nodes as and when required. A good load balancing algorithm must support scalability [10].

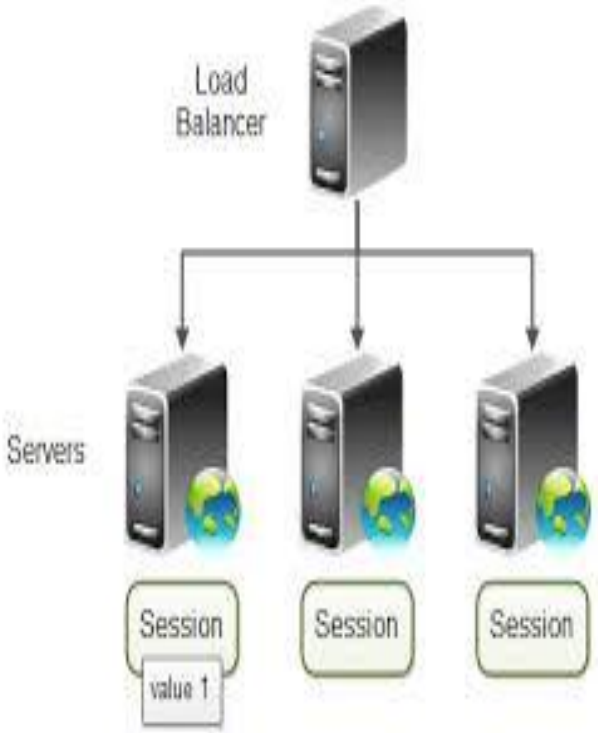

Figure 1.2 Load Balancer [12]

III. NEED OF LOAD BALANCING IN CLOUD

Load balancing is needed in many fields and one of its purposes is to accomplish green computing in clouds which is fulfilled by achieving namely two factors:

1. Reducing Energy Consumption: Load balancing aims at evenly distributing the workload among the various nodes that are engaged in performing tasks of the system, thereby reducing the amount of energy consumed [2].As the energy is reduced the system starts working more efficiently. The performance of the system is greatly enhanced.

2. Reducing Carbon Emission: Energy consumption helps in achieving carbon emission as they go hand in hand. The more the energy is consumed, higher is the carbon footprint and vice versa. So Load balancing helps in controlling the excess energy consumption so that the carbon emission is reduced which helps in achieving the green computing [2].

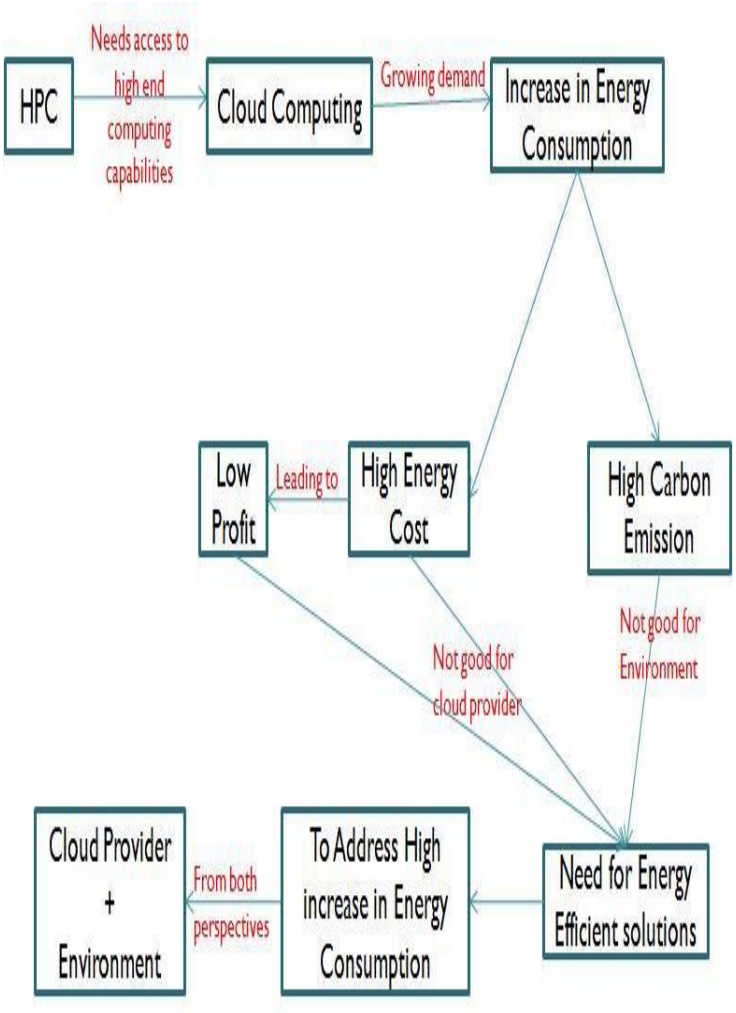

Figure 1.3 Green computing in Cloud [2]

IV. DIFFERENT LOAD BALANCING STRATEGIES

The main objective of load balancing is to reduce the effect of overloading and under loading of virtual machines by dividing the total load between the various numbers of machines. Basically there are two different load balancing strategies namely static load balancing and dynamic load balancing technique.

A) Static load balancing: In static load balancing algorithms prior knowledge of the system is known in advance which includes the characteristics of jobs, communication nodes and the communication network. The information is static and cannot be changed at the execution time. This type of algorithm has advantages in the terms of less overhead and easy implementation. However there are many problems associated with this kind of algorithm. One of them is that there are many applications which do not account for prediction of workload before the final program execution. Another issue is that these kinds of algorithms are not suited in distributed environment as everything is known in advance and which cannot be changed in between.[8]

B) Dynamic Load Balancing: These are those algorithms which do not follow the strict rules. They use current or recent load information when making distribution decisions [8]. Changes to the nodes can be made even while the execution of program. This kind of algorithm provides better performance than the static algorithms. The shortcomings of this algorithm are the time complexity and the communication overhead which occurred while exchanging information within the nodes [9].

Further these algorithms are classified into two approaches one is centralized and other is decentralized approach.

In centralized approach only one master node is there whose task is to assign the jobs to the various slave nodes. The slave nodes perform the job allotted by the master node.

The disadvantage of this node is that there is single point of failure in this approach. Here if anyhow the master node stops working then the whole system will get corrupted. Another approach is decentralized in which every node act as a participant in load balancing. The advantage of this approach is that it is more reliable and fault tolerant. The only shortcoming is the cost spends in communication delays [8].

V. CHALLENGES OF LOAD BALANCING IN CLOUD

A) Spatial Distribution of the Cloud Nodes

Some algorithms are designed to be valuable only for an intranet or closely located nodes where communication delays are negligible. However, it is quite a challenge to design a load balancing algorithm that can work for spatially distributed nodes. This is because of other factors which must be taken into account such as the speed of the network links among the nodes, the distances between the nodes involved in providing the service and the distance between the client and the task processing nodes. A way

needs to be developed to control load balancing mechanism among all the spatial distributed nodes while being able to effectively tolerate high delays [4].

2. Storage/ Replication

Full replication algorithms do not take efficient storage utilization into account. This is because of the same data that get stored in all replication nodes. Full replication algorithms often impose higher costs since more storage is needed. However, partial replication algorithms are capable of saving parts of the data sets in each node (with a certain level of overlap) based on the capabilities of each node such as processing power and capacity .[5]

This could lead to better utilization; still it increases the complexity of the load balancing algorithms as they attempt to take into account the availability of the data set's parts across the different cloud nodes [5].

3. Algorithm Complexity

Load balancing algorithms are favoured to be less complex in terms of operations and implementation. Higher the implementation complexity it would lead to a more complex process which could cause some negative performance issues.

Moreover, when the algorithms require more information and higher communication for control and monitoring, delays would cause more problems and the efficiency will eventually drop. Therefore, it is recommended that load balancing algorithms must be designed in the simplest possible forms [6].

4. Point of Failure

Collecting data about the different nodes must be designed in a way that avoids having a lone point of failure in the algorithm. Some of the algorithms (centralized algorithms) can provide efficient and effective mechanisms for solving the load balancing in a assertive pattern. However, they could have the issue of one controller for the whole system. In such cases, if the controller fails, then the whole system will also fail. So any Load balancing algorithm must be designed in order to overcome this challenge [7].

Distributed load balancing algorithms seem to provide a better approach than centralized load balancing algorithms, yet they are much more complicated and require more coordination and control to function correctly.

VI. SOME EXISTING LOAD BALANCING ALGORITHMS

- **Ant Colony Optimization:** Ant Colony Optimization (ACO) metaheuristic is inspired by the behavior of real ants finding the shortest path between their colonies and a source of food. This novel approach was introduced by Dorigo in 1992 in his Ph.D. thesis and was originally called ant system. While walking ants leave a chemical toxin substance called Pheromone on their way. The intensity of pheromone is greatly dependent on the no. of ants passing through and it can be dropped if the substance tends to evaporate [14].This algorithm works when the request is initiated by the tasks thereby ants start moving to search for food [15]. Pheromone substance intensity is helpful in locating food source to the ants [14].
- **PSO Based Particle Swarm Optimization:** Particle Swarm Optimization (PSO) is an evolutionary computational technique introduced by Kennedy and Eberhart [64] in 1995 motivated by social behavior of the particles. Each particle is allied with position and velocity and moves through a multi-dimensional search space. In each iteration, each particle adjusts its velocity based on its best position and the position of the best particle of the whole population. PSO combines local search methods with global search methods trying to balance exploration and exploitation. PSO has gained popularity due to its simplicity and its usefulness in broad range of applications with low computational cost [14].

CONCLUSION

Cloud computing is an extensive emerging technology in the computer world. Various issues exist in the era of computing. One of those issues is load balancing. Many algorithms are in existence for the load balancing of resources which are utilized in task mapping. There are various parameters associated with the working of load balancing algorithms which need to be controlled. So we believe there is wide scope of research in this field.

Our proposed taxonomy will help researcher to understand the concept of load balancing of tasks, need of load balancing, and several strategies associated with it and challenges need to be coped up with the load balancing.

REFERENCES

[1] T-Y., W-T. Lee, Y-S. Lin, Y-S. Lin, H-L. Chan and J-S. Huang, 2012 "Dynamic load balancing mechanism based on cloud storage" in proc. Computing, Communications and Applications Conference (ComComAp), IEEE, pp: 102-106, January.

[2] Nidhi Jain Kansal, Inderveer Chana "Cloud Load Balancing Techniques: A Step Towards Green Computing" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012 ISSN (Online): 1694-0814.

[3] Klaithem Al Nuaimi, Nader Mohamed, Mariam Al Nuaimi and Jameela Al-Jaroodi," A Survey of Load Balancing in Cloud Computing: Challenges and Algorithms" 2012 IEEE Second Symposium on Network Cloud Computing and Applications

[4] Buyya R., R. Ranjan and RN. Calheiros, "InterCloud: Utility-oriented federation of cloud computing environments for scaling of application services," in proc. 10th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), Busan, South Korea, 2010.

[5] 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.

[6] Grosu, D., A.T. Chronopoulos and M. Leung, "Cooperative load balancing in distributed systems," in Concurrency and Computation: Practice and Experience, Vol. 20, No. 16, pp: 1953-1976, 2008.

[7] Ranjan, R., L. Zhao, X. Wu, A. Liu, A. Quiroz and M. Parashar, "Peerto- peer cloud provisioning: Service discovery and load-balancing," in Cloud Computing - Principles, Systems and Applications, pp: 195-217, 2010.

[8] S. Dandamundi, Sensitivity Evaluation of Dynamic Load Sharing in Distributed Systems, Technical Report TR 97-12, Carleton University, Ottawa, Canada

[9] K. Lu, R. Subrata, A. Zomaya, An Efficient Load Balancing Algorithm for Heterogeneous Grid Systems considering Desirability of Grid Sites, 25th IEEE International Conference on performance, Computing and Communication, April 2006.

[10] Amandeep1,Vandana Yadav2, Faz Mohammad3, "Different Strategies for Load Balancing in Cloud Computing Environment: a critical Study" International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 3 Issue 1, April 2014

[11] Anurag Jain, Dr. Rajneesh Kumar, A taxonomy of Cloud Computing, International Journal of Scientific and Research Publications, Volume 4, Issue 7, July 2014 1 ISSN 2250-3153

[12]http://en.wikipedia.org/wiki/computing

[13]IBM, Fundamentals of Cloud Computing, Instructor Guide ERC 1.0, (November 2010) 17-230.

[14] Kalra M, Singh S, A review of metaheuristic scheduling techniques in cloud computing, Egyptian Informatics J (2015), http://dx.doi.org/10.1016/j.eij.2015.07.001

[15] Amandeep, Vandana Yadav, Faz Mohammad, Different Strategies for Load Balancing in Cloud Computing Environment: a critical Study, International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882
Volume 3 Issue 1, April 2014

BIOGRAPHIES

Anurag Jain is working as an Associate Professor in Computer Science Department, Chandigarh Engineering College, Landran, Mohali. He is in teaching field since 2003. He received his B. Tech. & M.Tech. degree from Kurukshetra University, Kurukshetra, India, in 2003 and 2009, respectively. He has guided 7 M. Tech. students. He has about 20 publications in International Journals and Conferences. Currently, he is pursuing Ph.D. from M, M, University Mullana Ambala in the area of cloud computing. His research interests are in the areas of load balancing & security in Cloud Computing. His email id is er.anuragjain14@gmail.com.

Divya Garg is working as a student in computer Science Department, Chandigarh Engineering College, Landran, Mohali. She has obtained her B.Tech. degree from Kurukshetra university, Kurukshetra. She is working in the area of cloud computing. Her interests are in the area of load balancing and security in cloud computing. Her email id is divyagarg1900@gmail.com .

