Review of Load Balancing Algorithms in Cloud Computing Paradigm

Durgesh Patel[1] Mr. Narendra Rathore[2]
[1]M. E. Scholar, [2]Assistant Professor
[1,2]CSE Department, SVITS Indore
[1,2]RGPV Bhopal, M.P., India

Abstract--- In modern days cloud computing is one of the greatest platform which provides storage of data in very lower cost and available for all time over the internet. But the cloud computing has more critical issue like security, load balancing and fault tolerance ability. In this paper we are focusing on Load Balancing approach. The 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 is required to handle the load when one node is overloaded. When the node is overloaded at that time load is distributed over the other ideal nodes. Many load balancing algorithms are available for load balancing like Static load balancing and Dynamic load balancing.

Keywords: Cloud Computing, virtualization, Load balancing.

I. INTRODUCTION

Cloud computing is an on demand service in which shared resources and other devices are provided according to the clients requirement at specific time. Cloud computing is a term which is generally used in case of Internet. The whole Internet world can be viewed as a cloud. The Capital and operational costs can be cut using cloud computing.

The Cloud computing [7,8,9] is a internet based network. Cloud is a collection of services. Cloud provides on demand services. The major services provided through cloud are: hardware service, software service, network service. Cloud computing is a modern field, which revolves around utility computing, service oriented architecture, internet, clients etc.

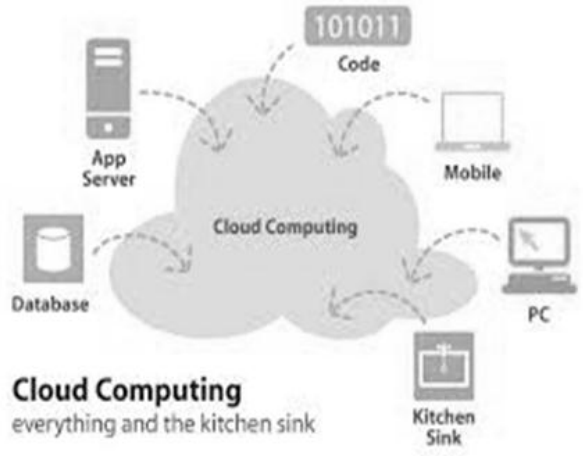

Fig. 1: Cloud computing Environment [10]

Now a days, cloud computing is the heart favorite topic to many researchers. It will become more popular in coming years as the reach of internet is increasing day by day.

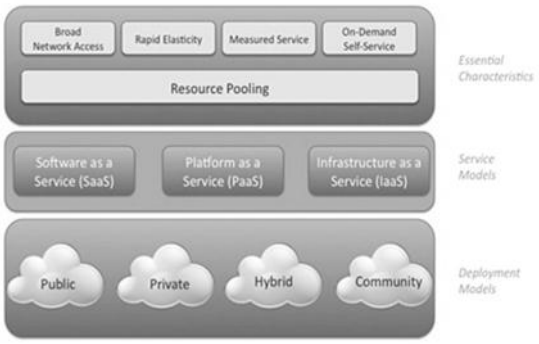

Fig. 2: Cloud Architecture [10]

Cloud computing has three basic models, which are Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS) . The Main advantages of cloud computing are: low cost, improved performance, infinite storage space etc.

Load balancing in cloud computing systems is really a challenge now. A distributed solution is required. As it is not always practically feasible or cost efficient to maintain one or more idle services just as to fulfill the required demands. All jobs can't be assigned to appropriate servers and clients individually for efficient load balancing as cloud is a very complex structure and components are present throughout a wide spread area.

II. LOAD BALANCING IN CLOUD COMPUTING

The load balancing is the process of distributing the load among various resources in any system. Therefore load need to be distributed over the resources in cloud-based architecture so that each resources does approximately the equal amount of task at any point of time. The basic need is to provide some techniques to balance requests to provide the solution of the application faster. All cloud vendors are based on automatic load balancing services, it allows clients to increase the number of CPUs or memories for their resources to scale with increased demands. These services are optional and depend on the clients business needs. So the load balancing serves two important needs, firstly to promote availability of Cloud resources and secondarily to promote performance [24].

In order to balance the resources it is important to recognize a few major goals of load balancing algorithms:

1) Cost effectiveness: first aim is to achieve an overall improvement in system performance at a reasonable cost.
2) Scalability and flexibility: distributed system in which the algorithm is implemented may change in size or

topology So the algorithm must be scalable and flexible enough to allow such changes to be handled easily.

3) Priority: scheduling of the resources or jobs need to be done on beforehand through the algorithm itself for better service to the important or high prioritized jobs in spite of equal service provision for all the jobs regardless of their origin.

III. LITERATURE SURVEY

Types of Load Balancing Algorithms:

A. Static Algorithms:

Static algorithms divide the traffic similarly between servers. Using this approach the traffic on the servers will be disdained easily and consequently it will make the situation more imperfectly. Round robin algorithm which divides the traffic equally is announced as round robin algorithm. There were lots of problems appeared in this algorithm. The weighted round robin was defined to improve the critical challenges associated with round robin. Each servers have been assigned a weight and according to the highest weight they received more connections. In the situation when all the weights are equal, servers will receive balanced traffic [11].

B. Dynamic Algorithms:

Dynamic algorithms designated proper weights on proper servers and by searching in whole network a lightest server preferred to balance the traffic. Selecting an appropriate server needed real time communication with the networks which will lead to extra traffic added on system. Comparison between these two algorithms although round robin algorithms based on simple rule, but more loads conceived on servers and thus imbalanced traffic discovered as a result [11].

Following load balancing techniques are currently prevalent in clouds.

The work done by A. Singh et al. [13] proposed a novel load balancing algorithm called VectorDot. This algorithm handles the hierarchical complexity of the datacenter and multidimensionality of resource loads across servers network switches and storage in an agile data center that has integrated server and storage virtualization technologies.

The work done by Stanojevic et al. [14] proposed a mechanism CARTON for cloud control that unifies the use of LB and DRL. The LB (Load Balancing) is used to equally distribute the jobs to different servers so that the associated costs can be minimized and DRL (Distributed Rate Limiting) is used to make sure that the resources are distributed in a way to keep a fair resource allocation.

Author Y. Zhao et al. [15] addressed the problem of intra-cloud load balancing amongst physical hosts by adaptive live migration of virtual machines. The load balancing model is designed and implemented to reduce virtual machines migration time by shared storage to balance load amongst servers according to their processor or IO usage.

Work done by V. Nae et al. [16] presented an event driven load balancing algorithm for real-time Massively Multiplayer Online Games (MMOG). The algorithm after receiving capacity events as input, also analysis its components in context of the resources and the global state of the game session, then generating the game session load balancing actions.

The J. Hu et al. [17] proposed a scheduling strategy on load balancing of VM resources that uses historical data and current state of the system. Proposed strategy achieves the best load balancing and reduced dynamic migration by using a genetic algorithm.

The A. Bhadani et al. [18] proposed a Central Load Balancing Policy for Virtual Machines (CLBVM) that balances the load evenly in a distributed virtual machine/cloud computing environment.

The LBVS H. Liu et al. [19] proposed a load balancing virtual storage strategy (LBVS) that provides a large scale net data storage model and Storage as a Service model based on Cloud Storage. The Storage virtualization is achieved using an architecture that is three-layered and load balancing is achieved using two load balancing modules. It helps in improving the efficiency.

The Y. Fang et al. [20] discussed a two-level task scheduling mechanism based on load balancing to meet dynamic requirements of users and obtain high resource utilization. Algorithm achieves load balancing by first mapping tasks to virtual machines and then virtual machines to host resources thereby improving the task response time, and resource utilization also overall performance of the cloud computing environment.

Author M. Randles et al. [21] investigated a decentralized honey bee based load balancing technique that is a nature inspired algorithm for self-organization. Algorithm achieves global load balancing through local server actions. Performance of the system is enhanced with increased system diversity but throughput is not increased with an increase in system size. This is best suited for the conditions where the diverse population of service types is required.

The work done by M. Randles et al. [21] 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.

Author M. Randles et al. [21] investigated a self-aggregation load balancing technique that is a self-aggregation algorithm to optimize job assignments by connecting similar services using local re-wiring. Overall performance of the system is enhanced with high resources thereby in-creasing the throughput by using these resources effectively.

The Z. Zhang et al. [22] proposed a load balancing mechanism based on ant colony and complex network theory (ACCLB) in an open cloud computing federation. Proposed algorithm uses small-world and scale-free characteristics of a complex network to achieve better load balancing. Proposed technique overcomes heterogeneity is adaptive to dynamic environments and has good scalability hence helps in improving the performance of the system.

Author S.-C. Wang et al. [23] proposed a two-phase scheduling algorithm that combines OLB (Opportunistic Load Balancing) and LBMM (Load Balance Min-Min) scheduling algorithms to utilize better executing efficiency and maintain the load balancing of the system. This OLB scheduling algorithm keeps every node in working state to achieve the goal of load balance and LBMM scheduling algorithm is utilized to minimize the execution time of each task on the node thereby minimizing the overall completion

time.

Author H. Mehta et al. [24] Proposed a new content aware load balancing policy named as work-load and client aware policy (WCAP). Proposed work uses a parameter named as USP to specify the unique and special property of the requests as well as computing nodes. The USP helps the scheduler to decide the best suitable node for processing the requests.

Author Y. Lua et al. [25] proposed a Join-Idle-Queue load balancing algorithm for dynamically scalable web services. Work provides large-scale load balancing with distributed dispatchers by, first load balancing idle processors across dispatchers for the availability of idle processors at each dispatcher and then, assigning jobs to processors to reduce average queue length at each processor.

IV. CONCLUSION

In this paper we have proposed a survey of load balancing methods. In cloud computing load balancing is one of the main issue. When client is requesting for service it should be available to the client. When any node is overloaded with job at that time load balancer has to set that load on another free node. Therefore load balancing is necessary in cloud computing. so in this paper we have discussed all the existing techniques for Load balancing.

REFERENCES

[I] John Harauz, Lorti M. Kaufinan. Bruce Potter, "Data Security in the World of Cloud Computing", IEEE Security & Privacy, Co published by the IEEE Computer and Reliability Societies, July/August 2009.

[2] National Institute of Standards and Technology-Computer Security Resource Center -www.csrc.nist.gov

[3] http://en.wikipedia.org/wiki/Cloud_computing.

[4] Yashpalsinh Jadeja and Kirit Modi, "Cloud Computing - Concepts, Architecture and Challenges", International Conference on Computing, Electronics and Electrical Technologies [ICCEET], IEEE-2012.

[5] Samerjeet kaur, "Cryptography and Encryption in Cloud Computing", VSRD International Journal of Computer Science and Information Technology, VSRDIJCSIT, Vol. 2 (3), 2012.

[6] Ramgovind S, Eloff MM, Smith E, 'The management of security in cloud computing", IEEE – 2010.

[7] Aderemi A. Atayero and Oluwaseyi Feyisetan," Security Issues in Cloud Computing: The Potentials of Homomorphic Encryption" Journal of Emerging Trends in Computing and Information Sciences, VOL. 2, NO. 10, October 2011. [8] Turban, E; King, D; Lee, J; Viehland," Chapter 19: Building E-Commerce Applications and Infrastructure". Electronic Commerce A Managerial Perspective. pp. 27, 2008.

[9] J. Kruskall and M. Liberman."The Symmetric Time Warping Problem: From Continuous to Discrete. In Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison", pp. 125-161, Addison-Wesley Publishing Co., 1983.

[10] Mr. Nitin S. More, Mrs. Swapnaja R. Hiray and Mrs. Smita Shukla Patel," Load Balancing and Resource Monitoring in Cloud", International Journal of Advances in Computing and Information Researches ISSN: 22774068, Volume 1– No.2, April 2012.

[11] R. X. T. and X. F. Z,"A Load Balancing Strategy Based on the Combination of Static and Dynamic, in Database Technology and Applications (DBTA)",2nd International Workshop,2010.

[12] David Escalante and Andrew J. Korty, "Cloud Services: Policy and Assessment", EDUCAUSE Review, vol. 46, no. 4 (July/August 2011).

[13] Singh A., Korupolu M. and Mohapatra D., ACM/IEEE conference on Supercomputing, 2008.

[14] Stanojevic R. and Shorten R., IEEE ICC, 1-6, 2009.

[15] Zhao Y. and Huang W., 5th International Joint Conference on INC, IMS and IDC, 170-175, 2009.

[16] Nae V., Prodan R. and Fahringer T., 11th IEEE/ACM International Conference on Grid Computing (Grid), 9-17, 2010.

[17] Hu J., Gu J., Sun G. and Zhao T., 3rd International Symposium on Parallel Architectures, Algorithms and Programming, 89-96, 2010.

[18] Bhadani A. and Chaudhary S., 3rd Annual ACM Bangalore Conference, 2010. [19] Liu H., Liu S., Meng X., Yang C. and Zhang Y.,International Conference on Service Sciences (ICSS), 257-262,2010.

[20] Fang Y., Wang F. and Ge J., Lecture Notes in Computer Science, 6318, 271-277, 2010.

[21] Randles M., Lamb D. and Taleb-Bendiab A., 24th International Conference on Advanced Information Networking and Applications Workshops, 551-556, 2010.

[22] Zhang Z. and Zhang X, 2nd International Conference on Industrial Mechatronics and Automation, 240-243, 2011.

[23] Wang S., Yan K., Liao W. and Wang S, 3rd International Conference on Computer Science and Information Technol-ogy, 108-113, 2010.

[24] Mehta H., Kanungo P. and Chandwani M., International Conference Workshop on Emerging Trends in Technology, 370-375, 2011.

[25] Lua Y., Xiea Q., Kliotb G., Gellerb A., Larusb J. R. and Green-ber A,"Int. Journal on Performance evaluation",2011.

