Load Balancing in Cloud Computing: A Review

Shikha Gupta, Suman Sanghwan

***Abstract*— A rapid growth in the development of clouds and its management through cloud computing has accelerated the research in this field. After seeing the growth it can say that the future of internet technology is totally based on cloud computing. It provide "as a service" on demand of user. It can be a software, platforms or infrastructure. The cloud owner's relationship with the consumer highly depends upon how efficiently the consumers are able to use the cloud resources, which in turn depend upon the effective cloud management. Many resources, big data and high demand, may deteriorate the service due to heavy loading of the server. This calls for the balance load on server by distributing the task to the appropriate node in the server. This paper presents a critical review and comparison of the existing techniques for load balancing.**

***Index Terms*—cloud computing, load balancing, data center, cluster.**

I. Introduction

In a current scenario of IT industry cloud computing has become an emerging technology. It is growing so fast and with the use of cloud computing, computing become the 5th utility of the daily needs after water, electricity, gas and telephony [1]. Cloud computing is more than a simple virtualization, even virtual computers are only the component of cloud computing. To understand cloud computing it is necessary to understand the technologies that are involved in cloud computing. While using cloud computing it tries to separate the application from the operating system and operating system from the hardware that runs everything. If hardware dies the operating system and application keeps running. Underlying concept of cloud computing is based on 'web application', 'clustering', 'terminal services servers', 'application servers', virtualization' and 'hosted instance'. When numbers of computers (node) are connected together and forming a cluster in the cloud, it may b possible that some node become overloaded because of the random request of services by the clients. Because of the unbalanced cluster the performance of cloud will get worst. This condition is raising a high demand of load balancers or effective load balancing techniques. Effective load balancing results in minimizing resource consumption, implementing fail-over, enabling scalability, avoiding bottlenecks and over-provisioning [2]. One more condition can arise when load balancer need to balance the traffic i.e. when an application requests to be uploaded to the cloud.

The paper presents a survey on the existing load balancing algorithms of Cloud Computing environments. An overview of these algorithms is given. The rest of this paper is organized as follows. Section II discuss about the cloud computing. Section III discusses the need of load balancing. After that, some technical aspects of load balancing is discuss in Section IV. In section V overview of load balancing algorithms is given. Then conclude the paper and show possible areas of enhancement and future plan of improving load balancing algorithms in Section VI.

II. Cloud Computing

Today in our daily life we are using cloud computing. While uploading (storing) the images (data) in social networking site the cloud computing being used. First thing in underlying concept of cloud computing is 'web application' simplest form of cloud computing. For example instead of installing Microsoft office at home computer the use of Google docx is a use of cloud computing. This can save the data while home computer gets crashed too. 'Cluster' in cloud computing can be create by connecting the various nodes. This uses the concept of virtualization. If one node get failed, can get the data of that node from any of the other node in the cluster. If any node in the cluster getting too much traffic because of the simultaneous request of the end user the concept of load balancing works here. There will b a load balancer connected to the cluster it realizes that any node in a cluster getting too much traffic it route the coming request to the another node in the cluster. Another technology that works in cloud computing is 'terminal services'. This works on the ancient technology of computer science where the main frame computer is used which were connected to a dumb terminals, here in cloud computing main frame is replaced by terminal service servers and dumb terminals are replaced by thin clients. Thin clients can be either a hardware or software anything. It can be a normal Mac computer or a windows computer with terminal services client installed, by clicking on terminal service icon it can directly connect to the terminal service server. All the job done by the thin client is actually happening at terminal service servers. Application servers are installed in a terminal services server. It restricts a thin client to access the terminal service server. It allows the thin client to access only the application for which it has permission to access. Now while using application server all the job will happen at this application server not at the terminal server.

When talk about cloud computing the first thing comes in mind is 'virtualization' on which the cloud computing

***Manuscript received May, 2015*.**

***Shikha Gupta**, Dept. of Computer Science and Engineering, DCRUST, Murthal, Murthal (Haryana), India, 9717383719*

***Suman Sangwan**, Dept. of Computer Science and Engineering, DCRUST, Murthal, (Haryana), India,*

ISSN: 2278 – 7798

rely. VMware is an example of client installed virtualization software. With the help of virtualization separation of the operating system from the hardware which runs the operating system can perform, means it gives a power to migrate operating system with the applications from one piece of hardware to another piece of hardware and everything remains intact. Hypervisor is another way to use virtualization. ESxi is an example of hypervisor, now with this hypervisor there is a need of management software for example VSphare.

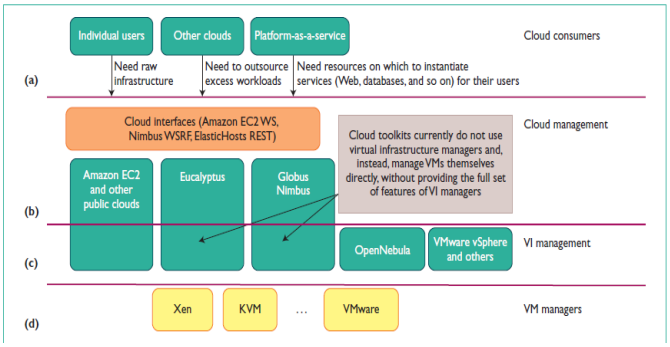

Fig 1. The cloud ecosystem for building private clouds [3].

(a) Cloud consumers need flexible infrastructure on demand. (b) Cloud management provides remote and secure interfaces for creating, controlling, and monitoring virtualized resources on an infrastructure-as-a-service cloud.
(c) Virtual infrastructure (VI) management provides primitives to schedule and manage VMs across multiple physical hosts.
(d) VM managers provide simple primitives (start, stop, suspend) to manage VMs on a single host.

There are three stakeholders in cloud computing [4].

1. End users in cloud computing-- A client of cloud who wants to use the services provided by the cloud that can be an end user in cloud computing. Services provided by the cloud is 'SaaS' (software as a services) in which there is no need to buy a software. Client can use the software on the basis of pay-per-use. SaaS is sometimes designated to as "on-demand software" and is commonly priced on a pay-per-use backbone. SaaS providers generally price applications using a subscription fee. GoogleApps and salesForce.com is an example of 'SaaS'. 'PaaS' (platform as a service) includes instead of buying platforms like database, web server, operating system, it can be used in the form of cloud on pay-per-use bases. Microsoft azure is an example of 'PaaS' 'IaaS' (infrastructure as a service) provides a virtual-machine, virtual storage, disk image library, virtual infrastructure , raw block storage, and file or object storage. Amazon EC2 is most known example of 'IaaS' [5].
2. Cloud provider in cloud computing— cloud provider who provide the services to the end user. They can offer a public cloud, private cloud and hybrid cloud on the bases of user request.

Public clouds are used by individuals or an organization based upon their requirements and necessities. They offer greatest level of efficiency in shared resources. There is security issue in using public cloud. They are more vulnerable than private clouds. Amazon web services, Google Compute Engine, Microsoft Azure, HP cloud are some of the public clouds. A hybrid cloud is a combination of public and private cloud [4][5].

3. Cloud developer in cloud computing— It bridge the gap between the client and cloud provider. The whole responsibility to develop the cloud is of cloud developer. It develops the cloud for cloud provider who provide it to the client.

III. Bases of Cloud Computing: Technical Aspects or Challenges

There are a number of technical challenges in cloud computing that need to be tackled before these benefits can be fully realized, which include infrastructure, load balancing, security and privacy in cloud computing, etc. Among them, load-balancing is a necessary mechanism to increase the service level agreement (SLA) and better uses of the resources.

1. Infrastructure- cloud provider has to manage all the hardware and network to provide the better services to the end user. if problem in infrastructure, that raise the issues in providing a services like SaaS and the cluster in cloud may get unbalanced due to poor infrastructure. This leads to a poor QoS [6].
2. Load balancing in cloud computing-- Major factor need to tackle in cloud computing is load balancing, many factors like poor infrastructure, bad traffic management, network reliability leads to unbalanced cluster. In small networks it can be negligible but in complex network, to provide better services all these are major factor to take care while designing algorithm for complex network [6].
3. Security and privacy in cloud computing— End users stores there data on the bases of security and privacy in cloud but due to many reasons like movement of data and application on network, loss of control on data, attacks on data, etc. the security may get effected. To recognize this issue is major challenge in cloud computing [7].
4. Trust in cloud computing-- When a client or end user request a service from cloud, there is a service level agreement needs to sign or needs to agree on the terms and condition of the cloud provider. All this is totally depends on trust of client on the cloud provider. Trust is an extended form of security and privacy. Two type of trust is defined in [7]. 1) hard trust (security-oriented) based on validity, encoding and security in 2) soft trust (non-security oriented) based on human psychology, loyalty to trade mark (brand loyalty) and user-friendliness

ISSN: 2278 – 7798

5. Ensuring data portability and interoperability-- There must be data portability in cloud computing, like the ability to change vendors in the future, agencies may attempt to avoid platforms or technologies that "lock" customers into a particular product.

IV. NEED FOR LOAD BALANCING

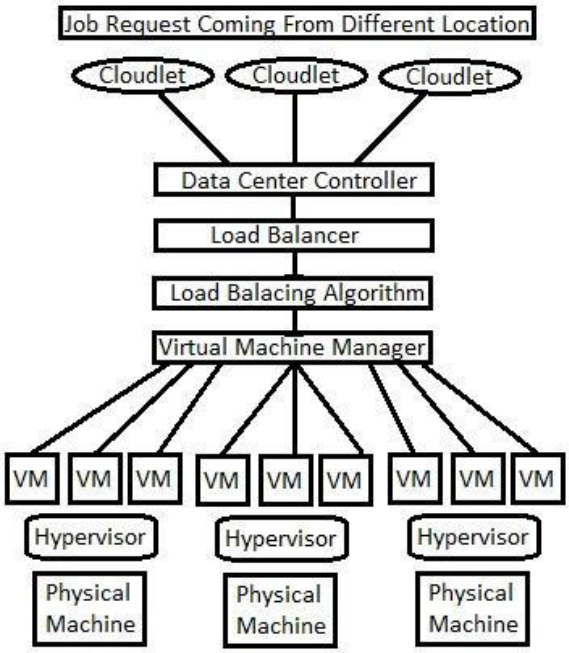

Fig 2. Structure of cloud computing environment [8].

The main aim of load balancing is to distribute the traffic among the node equally in the cluster for the better performance of network.

The aim of load balancing is as follows:

1. To enhance the surety of services to the consumer.
2. To enhance the user satisfaction.
3. To increase utilization of resource.
4. To reduce the execution time and waiting time of task coming from different location [8].
5. To make service performance better.
6. Maintain cluster stability.
7. Build a system that can tolerate the faults.
8. Reconcile future modification.

V. EXISTING LOAD BALANCING ALGORITHMS: A REVIEW

As an increased demand of the resources of cloud computing, load balancing is the usual problem to be faced. Various load balancing algorithms have been designed by various researchers. Load balancing algorithm can be categorized in two way static load balancing and dynamic load balancing.

In static load balancing we do not consider the dynamic changes in nodes of cluster during run time. It processed the node on the bases of prior information. While in dynamic load balancing we consider the chances in node information during runtime. It keeps the information of node up-to-date.

Static load balancing algorithms

In [8] author presents static load balancing algorithm i.e. load balancing min- min (LBMM) [8], in which the request waiting in a queue having minimum completion time allocate first to the node. Request having maximum waiting time have to wait in a queue until all the request get allocated to the node. Bottleneck of the request is the major issue in this algorithm. Well suited for the request having small completion time.

Opportunistic load balancing (OLB) is presented in [9] (very slow static load balancing algorithm). It believes only in keeping the node busy by assigning them to the request from the user. It does not take in to account the completion time of the node. When a node is processing one request it will not assign any task to it until it gets free from that task. It creates the congestion in the requesting queue and the request has to wait for a longer time in a queue for the node to get free.

Combination of LBMM and OLB is known as two phase load balancing. It tries to keep all the node busy on the bases of minimum completion time of the request.

Round Robin is illustrated in [10]. Based on the round robin scheduling algorithm of operating system a time slice is given to each node in a cluster. On the bases of this time slice the cloud provider provide the resources to the end user or client on its request for the service.

In [11] CLBDM (Central load balancing decision model) is given which is an extended form of round robin algorithm. In this external module is introduced which is connected to the nodes, load balancer application server etc. It calculate the time a node is spending with the client in sending and receiving the data with the help of a sensor placed in application layer. If it exceeds the time calculated by the sensor it moves the traffic to the other node using regular round robin.

Adaptive resource allocation is given in [12]. ARA algorithm is used to improve the decision making process of resource allocation and improve the system performance. It uses the best of greedy and random approaches. If there is request of a resource from an end user and the number of available resources are N, then ARA selects K resources from the available resources and randomly allocate one resource to the end user. The value of K should be decided on the bases of traffic of incoming jobs in the network. At the time of traffic the random behaviour of load balancer will work and the value of K should be equal to N. At the time of no traffic the greedy behaviour of load balancer will work and the value of K should b near to 1. In any other situation the value of K can be between 1 to N.

Dynamic load balancing algorithms

Honey bee forging behavior in [13] presents the algorithm which is inspired from the foraging behavior of honey bees. bees sent to search the suitable source of food is called a forager bees. When they found the source, they returned to the hive and advertise it in the form of "waggle dance". Found source is acceptable or not is decided on the bases of quantity and quality of nectar the bees harvested and

ISSN: 2278 – 7798

the distance of source from the hive. Now honey bees follow the forger bee to go to the source to harvest the food. After collecting the food they return to the hive and the remaining quantity of the food available at the source is shown in the form of "waggle dance", to decide that the remaining bees should sent to the same source or to search the new suitable source of food. In load balancing on the oscillation of demand for services by the end user.

In Biased random sampling presents in [13] says a virtual graph is used to represent the cluster in biased random sampling. Each node or computer of cluster is used as a virtual node of the virtual graph. The no. of free resources at the node represents the in-degree of the virtual node. This graph gives the current status of the network. The node having in-degree one, load balancer can allocate a task to the node. As the job is allocated to the node, the in-degree of a virtual node is decremented by one. As the task is completed it again increment it in-degree by one. This increment and decrement process is done by a random sampling. In virtual graph the in-degree of a node automatically become the out degree of another node which is a randomly selected node. By a sampling walk the balancer select the node to allocate the job. From a specific node it starts and move to randomly chosen neighbor node, at the node it stop the load is allocated to that node. The efficiency of the load distribution can be increased by increasing the walk length. Consider walk length is increased by *w*, the threshold value of walk length for the w will be log n where n is the size of the network. If the walk length of a node is equal to or greater than the threshold then the node is referred as a executing node. If it is less than the value of w is incremented and move to a next neighbor node. When the allocated job to the executing node has completed, the result of the allocation is shown a new edge from the initiating node to the executing node.

In [13] Active clustering is illustrated in which there is an initiator node and a matchmaker node in a network. Matchmaker node groups the similar type of node together. An initiator node randomly choose its neighbour as a matchmaker (matchmaker should not of its type). Now the matchmaker choose a node from its neighbor, if it is like a initial node then it connects the node from the initial node and remove the connection between itself and initiator node. Otherwise chosen neighbour of matchmaker become matchmaker. The process repeat until the entire node gets connected to its similar node. It gives the better utilization of resources, which results in increased system throughput.

The algorithm is inspired from the behavior of ants searching for their food is Ant colony optimization presents in [14]. Even the blind ants can reach till the food source with the help of the hint which the leading ants left for them. In cloud computing researchers uses this phenomenon to balance he node among the network. Head node is chosen in the cluster on the bases of degree of the node. The node having highest degree is elected to be a head node. Head node can b treated as a nest of ants from where they can go in various direction to search their food. Ants start their move from the head node. As it reached at the next node it checks whether the node is under loaded or over loaded, and move to next node. Again it checks that the node is under loaded or over loaded. This movement of ant is forward movement. If the previous node was over loaded and the current node is under loaded, it will go back to the previous node transfer the load from over loaded node to under loaded node or vice versa. This movement of ant is backward movement. The load of the network can equally distribute among the entire node in the network. In this algorithm a table is updated at each movement of ant to keep the network information up-to-date.

The extended form of ACO 'Ant colony and complex network theory' is given in [15]. Similar to ACO but a new feature is added to the algorithm is 'suicide' to reduce the congestion in the network. In this algorithm, after the completion of the job the ant commit suicide to reduce the backward movement in the network and all the procedure remain same as in ACO. This can work better with the complex networks.

In [16] author presents Index name server to keep the information about the network it uses distributed hash table, which reduces the duplication of data in databases. It applies the mathematical formula of distance and time to decide optimal path of a given weight in ad hoc network according to the paths' preference, and to figure out the performance of each node and the shortest path. The detail of INS is given in [16].

VI. Conclusion and Future Work

The underlying concept of cloud computing has been discussed with the various aspects of cloud computing. We have surveyed about various static and dynamic technology of load balancing in this paper. A large number of parameters and different types of soft computing techniques can be included in the future for the better utilization and needs of the user. In future some improvement in static ARA will solve the load balancing problem in dynamic network.

References

[1] R. Buyya, C. S. Yeo, S. Venugopal, 1. Broberg, I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Generation Computer Systems, vo1.25, pp. 599-616, June 2009.

[2] Nidhi Jain Kansal, Inderveer Chana, "Existing load balancing techniques in cloud computing: a systematic re-view," journel of information system and communication, ISSN: 0976-8742, E-ISSN: 0976-8750, Volume 3, Issue 1, 2012, pp- 87-91.

[3] 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.

[4] Mayanka Katyal, Atul Mishra, "A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment," International Journal of Distributed and Cloud Computing, Volume 1, Issue 2, December 2013.

[5] Alok singh, Vikas Kumar Tiwari, Dr. Bhupesh Gour, "A Survey on Load Balancing in Cloud Computing Using Soft Computing Technique's," International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 9, September 2014.

[6] Vikas Kumar, Shiva Prakash, "A Load Balancing Based Cloud Computing Techniques and Challenges," International Journal of scientific research and management, IJSRM,Volume 2, Issue 5, Pages815-824, 2014.

[7] Sajjad Hashemi, "cloud computing technology: security and trust challenges," International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 2, No 5, October 2013.

[8] Dharmesh Kashyap, Jaydeep Viradiya, " A Survey Of Various Load Balancing Algorithms In Cloud Computing," International Journal of Scientific & Technology Research, volume 3, issue 11, november 2014.

[9] 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 proc. 3rd International Conference on. Computer Science and Information Technology (ICCSIT), IEEE, Vol. 1, pp:108-113, July 2010.

ISSN: 2278 – 7798

[10] Nusrat Pasha, Dr. Amit Agarwal, 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, Issue 5, May 2014.

[11] 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.

[12] Jianzhe Tai, Juemin Zhang, Jun Li, Waleed Meleis, Ningfang Mi, "ARA: Adaptive Resource Allocation for Cloud Computing Environments under Bursty Workloads," IEEE, 2011

[13] Randles, M., D. Lamb and A. Taleb-Bendiab, "A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing," in Proc. IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA), Perth, Australia, April 2010.

[14] 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 proc. 14th International Conference on Computer Modelling and Simulation (UKSim), IEEE, pp: 3-8, March 2012

[15] Zhang, Z. and X. Zhang, "A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation." In proc. 2nd International Conference on. Industrial Mechatronics and Automation (ICIMA), IEEE, Vol. 2, pp:240-243, May 2010.

[16] [15] 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.

ISSN: 2278 – 7798