Cloud Computing based Load Balancing Architecture: A Study

Rajat[1], Dr. Sanjeev Kumar[2]
[1]M.Tech. Scholar, Department of CSE, GJUS&T, Hisar, Haryana, India.
[2]Assistant Professor, Department of CSE, GJUS&T, Hisar, Haryana, India.
Emails: [1]rajatman30@gmail.com, [2]sanjukhambra@yahoo.co.in

Abstract: Cloud Computing provides different services like network, software, application, server and more to the user, only internet connection is needed. In this user don't need to install the application and don't need local host or own gadgets to use the application rather than all the resources are shared by the users.Number of user that uses cloud computing is increasing exponentially as a result the traffic and the number of requests on the server also get increasing. Cloud computing has some critical issues like resource discovery, adaptation to non-critical failure, load balancing and security. To handle all the request and traffic in a better way load balancing is needed. Load balancing is one of the main issues in the cloud computing environment. Main motive of the load balancing is to enhance the usage of resources, deprecation reaction time, maximize resource utilization, enabling scalability, avoiding bottlenecks and avoiding the needless burden of any such assets. The main highlights of this document are the load balancing architectures available in the cloud environment. On the basis of distribution of load balancers in the framework there are three different architectures centralized, decentralized and hierarchical architecture that can be implemented for load balancing in the cloud environment. In this document firstly the load balancing is discussed in detail. After that all three load balancing architectures in cloud computing and their challenges are also reviewed.

Keywords –Load Balancer, Resource Pooling, Scheduling, Virtualization, Virtual server.

1. Introduction

In recent times cloud computing has turned out to be one in all the foremost discussed technology and has got a lot of considerations from the media and analyst due to the open doors it is putting forth.It is a web based model in which couple of independent abstraction such as infrastructural elements, application development and deployment environments, self contained software applicationsare delivered as a service to the user in the way as a user use the utilities like water and electricity (pay per use). To use these applications user only need internet connection and can use these any time via web browser[1]. Cloud Computing has been advancing over some stretch of time and many organizations are finding it fascinating to utilize. The entire Internet can be seen as a cloud.

NIST (National Institute of Standards and Technology) definition of cloud computing define the cloud as : "Cloud computing is a model for enabling ubiquitous, convenient,on-demand network access to a shared pool of configurable computing resources(e.g., networks, servers, storage, applications, and services) that can be rapidlyprovisioned and released with minimal management effort or service provider interaction"[2]. NIST definition of cloud clearly expresses that cloud computing helps in limiting company's consumption towards overseeing assets and also decrease the weight of keeping up programming or equipment by its client.

So by reducing cost and increasing flexibility cloud computing provides a number of benefits to the users. As a result a lot of companies and organisations get attracted towards cloud computing and the traffic or the data over the network is getting increased day by day. Both the cloud provider and the costumer want the best performance so the load on the network needs to be balanced. Load balancing is a process of balance or distributes the total load over all the nodes in order to get better performance [3].Load should be distributed in a way that at any point of time each assets should roughly have equal measure of work load or there should not be a situation that one node is overloaded while another has no work to do. In the casethat a node get failed and stop working, the system ought to ordinarily reload adjust the undertakings influenced to the inadequate asset so the user not get affected and still can use the cloud and can take the benefits [4], [5].

The main aim and motivation of this review paper is to tell about the entireLoad balancing architecture so that researcher can contribute in that area and can make better load balancing strategies. This paper start by rapidly depictingthe cloud computing that what it is and which services it provide to a customer. Thirdchapter focuses on the load balancing and its goals. The next part willtell about the different load balancing architecture and their advantage and disadvantage. Section 5proposes the conclusion and the final remarks and scope for future work.

2. Cloud Computing

Cloud Computing is anewly innovated technology. It is a model that offers computing over the web.A cloud computing service comprised ofextremely optimized virtual servers that providesnumerous software, hardware and data resources that can be used easily. Organisations can directly connect with the cloud and can utilize these services on pay per use premises. This helps corporations to stay away from the capital expenditure on additional on- premises infrastructure resources and instantly scale up and down according to requirement[6]. Cloud computing tries to separate application from the operating system and the hardware, so that there should be no dependency. So in cloud computing if operating system or hardware fails application services not get halt.

There is no doubt that cloud computing has a enormous numbers ofbenefits that an organisation can use.There are some fundamental qualities of cloud computing like virtualization, on demand self services, rapid elasticity, broad network access, resource pooling, measured service,. All kind of services that are provided by cloud vendors are separated into three areas -IaaS (Infrastructure as a service), PaaS (Platform as a Service) and SaaS (Software as a Service)[7]. All these services are provided and used in real time by using internet. IaaS include the services in which infrastructure like hardware, network, connectivity and storage is provided to the consumer whereas under PaaS readymade environment is provided as a service for managing, developing and testing the software application. SaaS refers to the category of service in which software are provided over the internet so neither need to install software nor have to pay cost to purchase the software and licence. All the services are used as a utility, in the same way as water or electricity is used.

There are three types of cloud in deployment model of cloud system. It includes Public, Private (On premises) and Hybrid cloud. Public cloud is same as internet based on the cloud computing standard model. Service provider uses the internet to make all the services available to the user. Services may be free or may be chargeable on the pay per use basis. Private or on premises clouds are owned by aindividual organization. It provides all the benefits that are provided by public cloud like flexibility, monitoring, automation and provisioning. It provides more security than cloud because implemented within a firewall. The mixture of both public and private is hybrid cloud. These two are mixed to take advantage of both and create more value[8].

Although cloud has many benefits but it is a developing technology so there are many issues and challenges that are still present in the area of cloud computing. There are many issue related to privacy, security and as the number of user is increasing exponentially some effective scheduling and load balancing techniques are needed.

3. Load Balancing

Load balancing is one of the key terms in the cloud computing which have a great effect on the performance of a framework, dependent on the measure of work allotted to a framework for a given period of time. It is the way toward adjusting the amount of workload among system assets for enhancing system performance and asset utilization [9]. Load balancing can minimize the reaction time, amplify the client's satisfaction, advances the system potency and can reduce the assignment dismissal [10]. In the cloud computing load will be CPU restrictions, limited memory capability, and network or postpone load.So overallload balancing try that there should not be a situation when one of the server or data center is under utilization and other get overloaded[11].

Load balancer has to follow some steps to perform the load balancing. These steps are:

- Receive incoming service requests from various clients.
- Calculate requested load size of the incoming load request from clients and build a request queue.
- Check the current load status of the serves in the server pool periodically using a server monitor daemon.
- Use a load balancing strategy/algorithm to select appropriate server.

The entire cloud provider supports automated load balancing. It permits consumer to extend the number of CPU's or memories for his or herassets to scale with expanded requests. Load balancing have many goals that have to be achieved in order to balance the resources[12],[13]. Main aims of load balancing are:

a) To maintain the system steadiness.
b) Adaptability and scalability: distributed framework in which the calculation or algorithm is executed could modify in size and topology.
c) Priority: Scheduling of the assets or jobs ought to be done before hand through the algorithmic program itself for providing appropriate services to the high priority jobs.
d) Cost effectiveness: To attain overall enhancement in system working at an inexpensive price.

All the providers use different types of load balancing strategies to achieve the various goals of load balancing. There are various metrics to measure that up to which extent load balancing goals achieved. Following are the metrics in existing load balancing techniques[14], [15]:

- **Throughput:** It is used to calculate the total number of tasks get completed in given amountof time. High value of throughput shows that load balancing is implemented properly.
- **Response time:** It is the measure of time taken by any load balancing strategy to react or respond in a framework. Low value of this time shows that load balancing is implemented in a good way.
- **Fault tolerance:** It is theability of a load balancing strategy to work in the situation even when some of the components stop working. A load balancing strategy should have good fault tolerance techniques.
- **Scalability:** Asystem with finite number ofresources should have ability to perform load balancing. An efficient algorithm should have optimized scalability.

- **Performance:** Performance refers to the overall proficiency or productivity of a framework. This must be enhanced at a sensible cost.
- **Resource utilization:** All the resources should be utilized in a good manner. It ought to be improved for a productive load balancing strategy.

4. Load Balancing Architecture

On the basis that how nodes are spatially distributed there are three type of load balancing architecture:

- Centralized load balancing.
- Decentralized load balancing.
- Hierarchical load balancing.

In **Centralized Load balancing** there is a central load balancer that is the main factor tomake decision about which algorithm (Static or Dynamic) should be used on the basis of global information about the state of system stored in it. As all the allocation and decision are made by a single node it reduces the overall time of load balancing. An incorporated Load balancer is placed in the client area that regularly collect the information from all the workstations and when a request arrives it is allotted to appropriated server on the basis of collected information. This method works well up to few thousand processors [16], [17].
Centralized method has a major problem that is the problem of scalability mostly in the case when machine has less memory. It also has some other limitations like single point of failure that is central node and it is very difficult to recover from that failure [18].

Decentralized Load Balancing techniquenot has a single node like centralized but it have number of load balancer to make the precise load balancing decisions. It provides awesome adaptability and versatility. However, in practice it faces problem like aging of load information and this can cause to make poor decisions at runtime. It happen because load balancing decision can only be made after getting all the workload information from the neighbour [19].
In this type of framework each processor shares its workload information with the neighbourhood processors, so small amount of memory is enough. Each load balancer in the system may use different technique for the load distribution. In this type of scenario chances of node failure is less. So none of the node get overloaded and computing environment become more reliable and fault tolerant[20].If single node gets failed then the whole network not gets much affected and if more than one node stops working it can also be tolerated. These type of networks are self-healing and self organising network so only require a little intervention from human operator[21].
.

Table 1: Comparison of load balancing architectures

Type of Architecture	Workload information Status	Advantage	Disadvantage
Centralized	Single node maintains global workload information.	Work well in small networks, Less overall time.	Nonscalable, Single point of failure, No longer fault tolerant.
Decentralized	Multiple nodes and each share information with neighbours.	Failure intensity of node is less, More reliable, fault tolerant.	Aging of load information, Poor load balancing on large machines.
Hierarchical	Nodes share information with child node.	Provide benefit centralized and decentralized.	Complex to implement, Less Fault tolerant.

In **Hierarchicalload balancing** load balancers are spatially distributed in a tree like structure, all upcoming requests are received by root node and then assignments are distributed to the load balancer at lower level. As in distributed method different load balancer can use different algorithm, same way in the hierarchical method load balancer at different level can use different algorithm. In this parent or root node is the main node that manage all the child node and distribute load to the child load balancer [22]. This method is combination of both centralized and the decentralized one, so provide the benefits of both.

Homogeneous and heterogeneous, both type of network can be implemented easily using Hierarchical architecture. It has some limitations also, like it is complex to implement this architecture. It also include additional overhead of deal with between the load balancers themselves [23].

5. Conclusion

Cloud Computing is most trending field of IT industry in recent times that uses the web to provide all the services like data, infrastructure and platform as a service. But still it is in developingand has many issues and challenges. Load balancing is one of the main areas out of all that still need a lot of research. Load balancing that is required to distribute the load among various processorsto get better performance.

This paper describe the concept of cloud computing and load balancing. Main gist of the paper is to discuss about different type of load balancing architecture that are present in cloud computing. Each one has its advantage and disadvantage. So it should be decided that which one is best for our scenario of cloud. In future researcher can do research that which architecture is best for which type of cloud and can propose a better architecture also. From this review it is concluded that as cloud is an emerging field and load balancing is main issue so there is a need of lot of research in this area to achieve better performance and business goals.

References

[1] L. Wang *et al.*, “Cloud computing: a perspective study,” *New Generation Computing*, vol. 28, no. 2, pp. 137–146, 2010.

[2] P. Mell, T. Grance, and others, “The NIST definition of cloud computing,” 2011.

[3] R. Kaur and P. Luthra, “Load balancing in cloud computing,” in *Proceedings of International Conference on Recent Trends in Information, Telecommunication and Computing, ITC*, pp. 374–381, 2014.

[4] M. Belkhouraf, A. Kartit, H. Ouahmane, H. K. Idrissi, Z. Kartit, and M. El Marraki, “A secured load balancing architecture for cloud computing based on multiple clusters,” in *Cloud Technologies and Applications (CloudTech), 2015 International Conference on*, pp. 1–6, 2015.

[5] N. J. Kansal and I. Chana, “Cloud load balancing techniques: A step towards green computing,” *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 1, pp. 238–246, 2012.

[6] S. M. Priya and B. Subramani, “A new approach for load balancing in cloud computing,” *International Journal of Engineering and Computer Science*, vol. 2, no. 5, pp. 1636–1640, 2013.

[7] B. P. Rimal, E. Choi, and I. Lumb, “A Taxonomy and Survey of Cloud Computing Systems,” presented at the Fifth International Joint Conference on INC, IMS and IDC, pp. 44–51, 2009.

[8] M. Nazir, “Cloud Computing: Overview & Current Research Challenges,” *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 8, no. 1, pp. 14–22, Dec. 2012.

[9] M. R. Mesbahi, M. Hashemi, and A. M. Rahmani, “Performance evaluation and analysis of load balancing algorithms in cloud computing environments,” in *Web Research (ICWR), 2016 Second International Conference*, pp. 145–151, 2016.

[10] A. Moghatadaeipour and R. Tavoli, “A new approach to improving load balancing for increasing fault tolerence and decreasing energy consumption in cloud computing,” in *2015 2nd internationl conference on knowledge-based on engineering and innovation*, Tehran, Iran, pp. 982–987, 2015.

[11] G. Joshi and S. K. Verma, “A Review on Load Balancing Approach in Cloud Computing,” *International Journal of Computer Applications*, vol. 119, no. 20, 2015.

[12] A. Garg, K. Patidar, G. K. Sexana, and M. Jain, “A Literature Review of Various Load Balancing techniques in Cloud Computing Environment,” *International Journal of Enhanced Research in Management & Computer Applications*, vol. 5, no. 2, pp. 11–14, Feb. 2016.

[13] T. Desai and J. Prajapati, “A survey of various load balancing techniques and challenges in cloud computing,” *International Journal of Scientific & Technology Research*, vol. 2, no. 11, pp. 158–161, 2013.

[14] B. Deepak, S. Shashikala, and V. Radhika, “Load Balancing Techniques in Cloud Computing: A Study,” *International Journal of Computer Applications*, pp. 1–4, 2014.

[15] N. J. Kansal and I. Chana, “Existing load balancing techniques in cloud computing: a systematic review,” *Journal of Information Systems and Communication*, vol. 3, no. 1, p. 87, 2012.

[16] A. Vig, R. S. Kushwah, and S. S. Kushwah, “An Efficient Distributed Approach for Load Balancing in Cloud Computing,” presented at the Computational Intelligence and Communication Networks (CICN), 2015 International Conference on, Jabalpur, India, pp. 751–755, 2015.

[17] J. C. Phillips, G. Zheng, S. Kumar, and L. V. Kalé, “NAMD: Biomolecular simulation on thousands of processors,” in *Supercomputing, ACM/IEEE 2002 Conference*, pp. 1–18, 2002.

[18] W. Zhu, C. Sun, and C. Shieh, “Comparing the performance differences between centralized load balancing methods,” in *Systems, Man, and Cybernetics, 1996., IEEE International Conference*, vol. 3, pp. 1830–1835, 1996.

[19] J. Yang, L. Ling, and H. Liu, “A Hierarchical Load Balancing Strategy Considering Communication Delay Overhead for Large Distributed Computing Systems,” *Mathematical Problems in Engineering*, vol. 2016, pp. 1–9, 2016.

[20] H. Mehta, P. Kanungo, and M. Chandwani, “Decentralized Content Aware Load Balancing Algorithm for Distributed Computing Environments,” presented at the International Conference and Workshop on Emerging Trends in Technology (ICWET) – TCET, Mumbai, India, pp. 370–375, 2011.

[21] G. Jackson, P. Keleher, and A. Sussman, “Decentralized Scheduling and Load Balancing for Parallel Programs,” presented at the 2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 324–333, 2014.

[22] M. Katyal and A. Mishra, “A comparative study of load balancing algorithms in cloud computing environment,” *International Journal of Distributed and Cloud Computing*, vol. 1, no. 2, pp. 5–14, Dec. 2013.

[23] A. Khiyaita, H. El Bakkali, M. Zbakh, and D. El Kettani, “Load balancing cloud computing: State of art,” in *Network Security and Systems (JNS2), 2012 National Days of*, 2012, pp. 106–109.