© 2016 IJSRSET | Volume 2 | Issue 4 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Load Balancing in Cloud Computing : A Survey

M. Ramya*, Dr. D. Ravindran

Department of Computer Science, St. Joseph's College, Trichy, Tamilnadu, India

ABSTRACT

"CLOUD…..CLOUD ….." become the buzzwords for the research field. The cloud has been provided by the service providers like Microsoft, Google, Rackspace, and IBM. These services are given to the users for pay-per-use concept. The term "Cloud computing" was first coined after the thought that services and applications are deployed into the internet "cloud". It is not the word something got instantly by one night; it has got by the ancestors of time sharing, web based systems, peer-to-peer networks. The cloud consumers can get all the benefits from the cloud service provider and whenever they don't want the services they can relieve from it. Then the providers will measure the service and produce the charges based on the consumption. The cloud service providers are having the characteristics like rapid elasticity, measure service, security, extended network access, resource pooling. The services given to users can be affected by many issues like security, service level agreement, and multi-level tenancy. This paper gives the review on the load balancing concepts, what are the previous methods to balance the traffic, issues in load balancing and its types. This paper also reviews the various techniques in load balancing cloud environment.

Keywords: Cloud Computing, Virtualization, Load Balancing, Static Algorithm, Dynamic Algorithm.

I. INTRODUCTION

Cloud computing is the common buzzword for the researchers and also for the technology people. It allows users to hit into an immense huge pool of shared computing resources such as servers, storage and network. These resources are given as a service to the end users allowing them to "plug into the cloud area" which is analogous to a utility grid [1]. The promise of the cloud provider is to free the users from the tedious complex process of managing and provisioning the computing resources to run the user applications.

The cloud brings several additional uses including: a pay-as-you-go cost aware model, at ease deployment of applications, elastic scalability, high availability of resources and more robust and secure infrastructure. The ability of the system to increase the system resources in cloud platform is called cloud scalability. A computing system is said to be scalable if it can able to handle increasing load simply by giving suggestion to use more computing resources. If for an instance, by adding more memory or by adding another CPU the above process can be implemented. Scale-out facility permits a system to handle larger workloads by adding more physical machines to the entire system. Systems that are elastically scalable are able to respond to changes in load by growing and shrinking their processing capacity on the demand. Ideally, at any given time, an application deployed in the cloud should be using exactly the amount of resources required to handle its load, even as this load fluctuates.

II. CLOUD COMPUTING

A cloud is a large spool of virtualized computer resources [2]. A cloud is an oversized pool of simply usable and accessible virtualized resources. Distributed computing takes us to a new and overwhelming technology called "Cloud Computing" used by both the academicians and industry people to store and retrieve the files and necessary documents. In this way, consumer consumptions are measured for services

according to how much they have actually used during the billing period.

A. **Characteristics**

Cloud computing has a variety of characteristics [3],

- Shared Arrangement — using the virtualized software model which enables the clients for sharing of physical servers, storage and networking facilities. Regardless of the cloud deployment models the cloud infrastructure strives to make almost all of the available resources across the number of users. By the quality of shared environment the users will make use of the resources globally anywhere.
- Dynamic Provisioning — allowing for the provisioning of the resources and services as per the current need. This allocation has been done automatically using software automation which enables the expansion and contraction of service capability when needed. This auto scaling needs to be done while maintaining high levels of reliability and security.
- Extensive Network Access — wishing to be accessed across the internet by a wide-range range of devices such as PCs, laptops and mobile devices by using standards-based APIs. The deployments of services in the cloud include everything for business applications to the latest application on the latest smartphones.
- Managed Metering — using the metering for billing and reporting the information, and to manage and to optimize the services which user wants.

B. Cloud Computing Models

The cloud computing models have been differentiated by their usage as [4]:

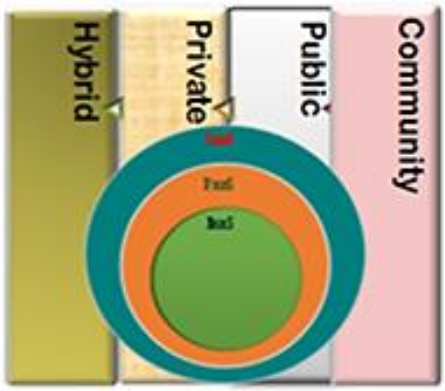

Figure 1. Cloud Models

1. Service Models

Once a cloud is established how and by what method its services are deployed in terms of business models can differ depending on requirements. The primary service models being deployed as

- Software as a Service (SaaS) — the consumers will use the application as per the need and they can purchase the software for the certain period of time. A good example of this is Salesforce.com. In this example, the necessary information for the interaction between the consumer and the service is hosted as part of the service in the cloud.
- Platform as a Service (PaaS) — the consumers will deploy their software and applications on the cloud through purchasing the access for the platforms. The consumer can't able to manage the two platforms under a single system.
- Infrastructure as a Service (IaaS) — the consumers will have control over and manage the systems in terms of the operating systems, applications, storage capacity and network connectivity, but they do not have control the cloud infrastructure. The concept of virtualization will be implemented in this service.

2. Deployment Models

The deploying cloud computing can differ depending on clients requirement and the succeeding four deployment models have identified with specific characteristics that support the needs of the services and users of the clouds.

- Private Cloud — the cloud infrastructure is deployed, maintained and operated for a specific

organization. This infrastructure operation may be built in-house or with a third party on the premises.

- Community Cloud — the cloud infrastructure is shared and operated among the specific organization those having the similar desires and requirements. This type of cloud may decrease the expenditure costs of the operations which are distributed among the organization sub-section.
- Public Cloud — the cloud infrastructure is open to all public on a commercial basis by the cloud service provider. This cloud model is developed and deploy with the very little financial outlay as compared to the other deployment options.
- Hybrid Cloud — the cloud infrastructure is developed by the combination of public and private cloud. The services based on the cloud type will be accessed. It has the ability to access the cloud services of any type of hybrid nature.

III. LOAD BALANCING

Load balancing is a process of assigning and distributing the total load to the individual separate nodes of the collective system. It facilitates the networks or resources to improve the response time of the job which is coming into the system with maximum throughput in the whole computing system. The term "Load Balancing" generally refers to transfer load from the overloaded process or resources to other under loaded process or resources [5]. The main problem to be considered in this issue is that how to choose the next node and when and where to transfer a load. It is done with the guide of load balancers where each incoming request is redirected and transparent to client who makes the request to the server. Based on predetermined parameters of the system such as availability or current load, the load balancer uses various scheduling algorithm to decide which server should handle and forwards the request on to the selected server. Virtualization technique has improved utilization of the computing system and system load balancing by enabling VM migration and has provided significant benefits for cloud computing. Based on the implementation method of load balancing algorithms can be classified in two ways.

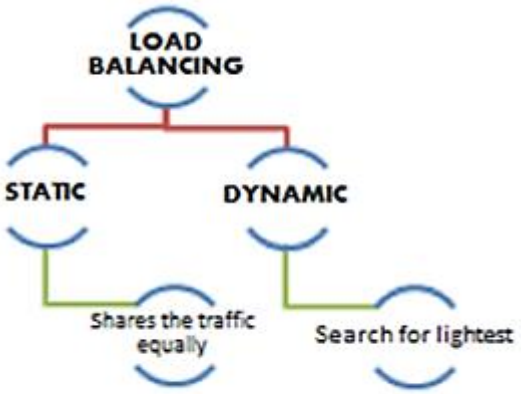

Figure 2. Types of load balancing

A. Goals of Load Balancing

The monitoring of resources with load balancing algorithm will increase the consumer satisfaction. The load balancing idea has been introduced with goals such as [5]

- Optimum utilization of resources
- High Throughput
- Short response time
- Avoiding bottlenecks

B. Performance metrics for load balancing algorithms

When the developer is going for any load balancing algorithm the following performance metrics has been considered in the system[5,6],

- Throughput: It has considered for the high performance of the system. In general terms it is defined as the amount of work completed in a specified period of time. It is the term used to calculate the total number of tasks those execution has been already finished.
- Response time: For better performance of the short response time is desired. In the distributed and parallel environment. In a distributed setting, the time it takes for a load balancing algorithm to start responding to a given instruction is called as response time for a better performance and short response time is desired. Generally, response time is

the amount of time taken to respond by a load balancing algorithm.

- Fault tolerance: It is needed for the optimal performance of the system. It is the ability of the load balancing algorithm to continue its operation even if some error occurs a high fault tolerance metric is mainly considered for optimal system performance. It is the ability of an algorithm to perform uniform load balancing in case of link failure to the systems.
- Scalability: It is the ability and capacity of the load balancing algorithm to manage the growing amount of workload by increasing the number of nodes based on the load factor. It is the ability to scale-up or scale-down agreeing to the requirements of the system.

- Resource utilization: it is the factor to which extent the resource is utilized for efficient performance. It measures the overall utilization of the computing resources using a particular load balancing algorithm.
- Forecasting accuracy: it is the degree of adoptability to which the expected output to which the one is calculated after calculation.
- Overhead Associated: It is the measure of amount of overhead included when implementing the algorithm for movement of tasks and inter-process communication.
- Migration Time: Migration Time is the essential time for which time required to migrate the jobs or resources from one node to another node. It should be minimized to enhance the overall performance of the cloud computing system.

IV. REVIEW OF LITERATURE

Mayanka Katyal et al. (2013) [7] start this paper by presenting the cloud computing, cluster, grid and previous history of cloud. The cloud perspectives like end users, providers, developer. In developing the cloud, these three persons will play the role. The each perspective has been explained with its requirements and issues. The authors are telling about the load balancing in the virtual systems and the cloud computing environment. Then they are describing about the resources allocation and task scheduling which is the important work of the load balancing execution. Further they made the brief comparison among the different types of load balancing algorithms. This gives the full details of the various algorithms in this issue.

Maram Mohammed Falatah et al. (2014) [8] author discussing about the definitions of cloud scalability. Scalability is the ability of the systems to do the works which are giving by the user in a fast manner. It has to take care of the many parameters of load balancing, resource allocation, and optimization. They had given the scalability levels and its performance considerations. This paper then presents the scaling approaches and further it gives the details about the study about the web application in cloud.

Subhadra Bose Shaw et al.(2014) [9] have discussed different proposed algorithms to resolve the issue of load balancing, task scheduling problem in cloud computing and also they have mentioned some of their shortcomings for further development. The load balancing algorithms are discussed with VM migration issues are also described briefly. They had made a brief comparison about the entire algorithm in the area of load balancing; its pros and cons are discussed.

Abhijit Aditya et al. (2015) [10] presents the basics of cloud computing like its characteristics, deployments models, service models. They are describing the each service delivery models characteristics, its vendor types their advantages and disadvantages. Then they describing about where the load problems are occurring in the system and so the challenges in keeping mind. Then it describes each and every types of algorithm in load balancing separately. Their properties, advantages, disadvantages are also described. They specially described about these algorithms based on the time factor.

Nadeem Shah et al. (2015)[11] describing this paper in order to understand the current challenges in cloud computing, mainly on cloud load balancing using the various static algorithms and finding breaks to bridge for more efficient static cloud load balancing in the future. The ideas suggested here are treated as new solution through which researchers allow to redesign best algorithms for good functionalities and improve the user experiences in simple cloud systems. This could assist small businesses that cannot afford infrastructure that supports complex & dynamic load balancing algorithms.

They gave the possible solutions for the current challenges in load balancing area.

Rajyashree et al. (2015) [12] describe the double threshold policy for describing the load balancer. They are concentrating mainly on the virtual machine level load balancing. They made the survey of five papers. They had proposed the mathematical calculation of virtual machine load on CPU, bandwidth, RAM. They are explaining about the virtual machine selection and based on that they are giving the lower and upper threshold work.

Po-Huei Liang et al.(2015)[13] presents a framework for global server using for load balancing of the web sites in a cloud with two hierarchical level load balancing model. The proposed framework is intended for adjusting an open-source load-balancing system and while the customers need more load balancers for increasing the availability, this framework allows the network service provider to deploy the load balancer in different data centers dynamically. This load balancing algorithm is suitable both for the hardware and software of the cloud computing architecture. Further they described the load balancing algorithms with the various cloud service providers along with its communication interface.

M.Kriushanth et al. (2015) [14] describing the auto scaling values and that setting dynamic threshold values in a cloud environment should utilize the available resources completely and prevents the physical server damage. It manipulates the provider to accommodate more users in a physical server and also reduces the cost of the service. In this paper, the authors elaborate their concept in the area to set a dynamic threshold value for the physical server, load balancer behavior identifier mechanism is proposed to generate the rule and provide the resources dynamically. Major drawback is that it can't be able to use in cloud data.

Akshada Bhujbal et al. (2015) [15] highly concentrates about the round robin and game theory algorithm for the high efficiency of the resource utilization. The load balancing code has been discussed and the architecture is introduced with main controller. They compute the load parameter like load status level, load degree. Though they cut down many parameters, security and some more efficiency has been increased by using various algorithms.

Radha Ramani Malladi (2015) [16] presents a concept of Cloud Computing along with load balancing. The main thing is considered in this paper is load balancing algorithm. There are various mentioned algorithms in cloud computing which consists of many factors like scalability, enhanced resource utilization, high performance and improved response time. Further this paper provides the insight about the policies, characteristics, goals, current state classification, need for load balancing. They have proposed a frame work for giving the new algorithm.

II. CONCLUSION

Cloud computing usage has been increasing day-by-day. The internet user will not know that simple email is implemented using cloud. It has vast varieties of cloud issues that can be improved by the researchers. The main issues to be considered under cloud are security, scalability, service level agreement, multi tenancy. The more scalability techniques have not been developed nowadays. Even though the researchers developed load balancing algorithms like round robin, weighted round robin, throttled algorithm and many combined algorithms there are many flaws in each algorithms.

III. REFERENCES

[1] Bhaskar Prasad Rimal, Eunmi Choi, Ian Lumb, "A Taxonomy and Survey of Cloud Computing Systems", IEEE, 2009.

[2] Sushil Kumar, Deepak Singh Rana, Sushil Chandra Dimri, "Fault Tolerance and Load Balancing algorithm in Cloud Computing: A survey", International Journal of Advanced Research in Computer and Communication Engineering, ISSN (Online) 2278-1021, ISSN (Print) 2319-5940, July 2015.

[3] Tejinder Sharma, Vijay Kumar Banga, "Efficient and Enhanced Algorithm in Cloud Computing", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, March 2013.

[4] Fahimeh Ramezani • Jie Lu, Farookh Khadeer Hussain, "Task-Based System Load Balancing in Cloud Computing Using Particle Swarm Optimization", Springer, 2014.

[5] Hardi S. Sanghavi, Dr. Tejas P. Patali, "Load Balancing Algorithms For The Cloud Computing Environment: A Review", Journal Of Information, Knowledge And Research In Computer Engineering, ISSN: 0975 – 6760, Oct 15.
[6] Amritpal Singh, "A Review of Existing Load Balancing Techniques in Cloud Computing", International Journal of Advanced Research in Computer Engineering & Technology, Volume 4 Issue 7, July 2015.
[7] 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.
[8] Maram Mohammed Falatah1, Omar Abdullah Batarfi, "Cloud Scalability Considerations", International Journal of Computer Science & Engineering Survey, Vol.5, No.4, August 2014.
[9] Subhadra Bose Shaw, Dr. A.K. Singh, "A Survey on Scheduling and Load Balancing Techniques in Cloud Computing Environment", International Conference on Computer and Communication Technology (ICCCT), ©2014 IEEE.
[10] Abhijit Aditya, Uddalak Chatterjee and Snehasis Gupta, "A Comparative Study of Different Static and Dynamic Load Balancing Algorithm in Cloud Computing with Special Emphasis on Time Factor", International Journal of Current Engineering and Technology, E-ISSN 2277 – 4106, P-ISSN 2347 – 5161 ©2015.
[11] Nadeem Shah, Mohammed Farik, "Static Load Balancing Algorithms In Cloud Computing: Challenges & Solutions", International Journal Of Scientific & Technology Research Volume 4, Issue 10, October 2015, ISSN 2277-8616.
[12] Rajyashree, Vineet Richhariya, "Double Threshold Based Load Balancing Approach by Using VM Migration for the Cloud Computing Environment", International Journal Of Engineering And Computer Science ISSN: 2319-7242, Volume 4 Issue 1 January 2015.
[13] Po-Huei Liang, Jiann-Min Yang, "Evaluation Of Two-Level Global Load Balancing Framework In Cloud Environment", International Journal of Computer Science & Information Technology, Vol 7, No 2, April 2015.
[14] M.Kriushanth, Dr. L. Arockiam, "Load Balancer Behavior Identifier (LoBBI) for Dynamic Threshold Based Auto-scaling in Cloud", International Conference on Computer Communication and Informatics, Jan. 08 – 10, 2015.
[15] Akshada Bhujbal, Prajakta Jakate, Manasi Wagh, Madhura Pise, Prof.M.V.Marathe, "Load Balancing Model in Cloud Computing", International Journal of Emerging Engineering Research and Technology Volume 3, Issue 2, February 2015, PP 1-6 ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online).
[16] Radha Ramani Malladi, "An Approach to Load Balancing In Cloud Computing", International Journal of Innovative Research in Science, Engineering and Technology, ISSN (Online): 2319 – 8753, Vol. 4, Issue 5, May 2015.