International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 1, January-February 2015 **ISSN 2278-6856**

Survey of Load Balancing and Scaling approaches in cloud

Priyanka P. Kukade[1] and Geetanjali Kale[2]

[1]Pune Institute of Computer Technology, Dhankawadi, Pune - 411 043

[2]Pune Institute of Computer Technology, Dhankawadi, Pune - 411 043

Abstract

In cloud, scaling is the ability to provide the services without lagging performance, in case of increased or decreased load. The system should be able to adapt to the customer request so as to increase resources or decrease the resources, so as to maintain the balance between performance and cost effectiveness. For scaling the application it is required to increase or decrease the application instances. While doing this load on physical machines also needs to be taken into consideration for maintaining a good performance. In this paper we have studied the existing load balancing algorithm and different approaches used for scaling in cloud computing.

Keywords: Cloud computing, Load Balancing. Virtual Machine, Resource Utilization.

1. INTRODUCTION

Cloud computing has emerged as a new computing model, where storage, network and computation is provided as a service to user, who can lease and release the service on demand [1]. It is a pay per use model, where you pay only for the resources you need .The National Institute of Standards and Technology [2] defines cloud computing as: "A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction".

Service delivery in Cloud Computing comprises three different service models, namely Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) as shown in the fig 1 [10]. Software as a Service (SaaS) is defined as: "Software that is deployed over the internet .With SaaS, a provider licenses an application to customers either as a service on demand, through a subscription, in a "pay-as-you-go" model, or (increasingly) at no charge when there is opportunity to generate revenue from streams other than the user, such as from advertisement or user list sales" [6].PaaS can be defined as a computing platform that allows the creation of web applications quickly and easily and without the complexity of buying and maintaining the software and infrastructure underneath it [6]. Infrastructure as a Service (IaaS) is a way of delivering Cloud Computing infrastructure – servers, storage, network and operating systems – as an on demand service [6].

With cloud computing it is possible to provide cloud services (software/infrastructure/platform) as per the requirement. Requirement can fluctuate as per customer needs and thus the provisioned resources can change. This feature of cloud computing is known as "elasticity", which has made cloud popular now a days. The below fig shows the general architecture of cloud network, where client request are handled by load balancer which starts the services on virtual machines as per load on each server.

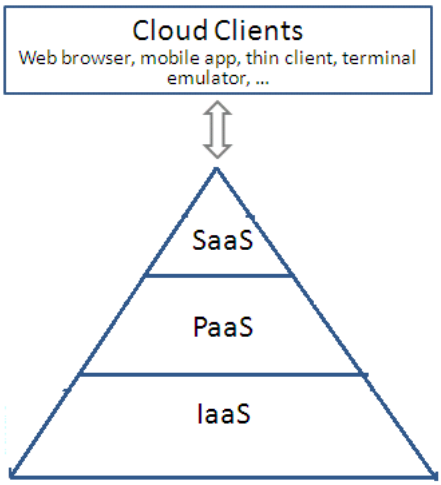

Figure 1 Service Model of Cloud

Elasticity is the key concept in cloud, which allows dynamically allocating and freeing resources as per request. This elasticity property is gained through scaling up and scaling down the services requested by the customer. Scaling should make it possible for
the system to deal with the increasing and decreasing of customer demands.

Even though the capacity of cloud is advocated to be infinite, the capacity of data center in real world is finite. When large number of service experiences their peak demand at the same time, the available resources become constrained. It is required to manage more instances of application when the demand increases and to scale down for less request for conserving energy. While scaling new instances of application need to be deployed on different web app server. Load on each server needs to be taken into consideration for this. Load distribution is balanced by migrating the load from the source nodes (which have surplus workload) to the comparatively lightly loaded

destination nodes.

The rest of the paper is organized with the sections as follows. As for scaling of application load balancing is also required as a part of scaling, we are presenting survey of the existing load balancing algorithm in Section 2. Section 3 gives the overview of different scaling approaches in cloud. Finally, we conclude the paper.

2. EXISTING LOAD BALANCING ALGORITHM

In cloud computing environment load balancing is required distribute the dynamic local workload evenly between all the nodes [5][8][9]. Load balancing is used to make sure that none of your existing resources are idle while others are being utilized. The below fig 2 shows the general architecture of cloud network, where client request are handled by load balancer which starts the services on virtual machines as per load on each server. The different load balancing strategies are discussed below.

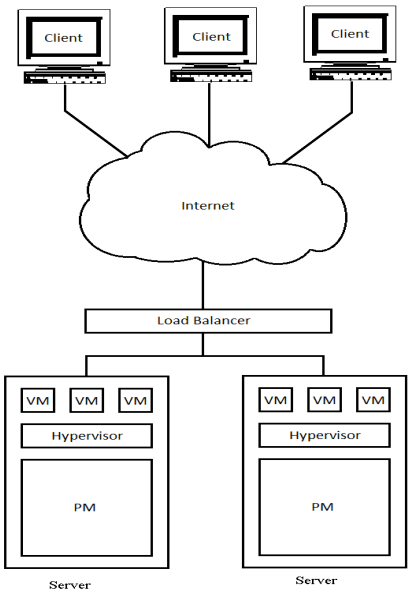

Figure 2 Cloud Architecture

Priya, S. Mohana, and B. Subramani has proposed an innovative load balancing technique Resource-Aware-Scheduling algorithm (RASA) [13]. This technique is a combination of Min-min and Max-min strategy. Here the virtual nodes are created first. Then response time of each VM is calculated and then accordingly least loaded node is found and client is given that node. The strategy is to apply Max-min if number of resources are even else Min-min strategy is used.

The throttled load balancing algorithm is explained by Sharma, Tejinder, and Vijay Kumar Banga [14]. Client request the VM to data center, which forwards it to load balancer. Load balancer maintains an index table of the available VM and their busy or available status. It then returns the ID of the first VM it encounters while scanning the list which can meet requirement to the data center. If the list is scanned completely and no VM is found it returns -1 to data center and data center queues the client request. When VM is available load balancer informs data center and it is allocated to client.

Shagufta Khan and Niresh Sharma have used modified approach of ant colony optimization [17]. In this approach ant travels in one direction at a time. It uses two directions of travelling-forward and backward. When the ant travels in forward direction it searches for overloaded node following the foraging pheromone and updates foraging trials. Same way, the ant travels in backward direction after encountering overloaded node by following the trailing pheromone and updates trailing pheromone trials in the path. It considers the minimum migration time of a node to find under loaded node.

Load balancing approach is described using honey bee foraging behaviour by Harshit Gupta, Kalicharan Sahu [19]. A fraction of honey bee called forager bees search for the food source. On finding the food source they return to beehive and perform waggle dance. Waggle dance gives indication of the quality and quantity of the food. Scout bees then follow forager bees to location and reap the food. Scout bees then return to the beehive and perform waggle dance to display the amount of food left. In this paper the tasks are to be send to the under loaded machine and like foraging bee the next tasks are also sent to that virtual machine till the machine gets overloaded as flower patches exploitation is done by scout bees. But their strategy considers minimum migration time factor and does not consider all QoS factors.

3. SCALING APPROACHES IN CLOUD

Elasticity property of cloud is gaining more attraction of people. Horizontal and vertical scaling are the ways by which scaling can be accomplished as shown in the fig 3. Horizontal scaling deals with adding or removing of resources that are of the same type. Vertical scaling deals with replacing the existing resource with the lower or higher capacity. Different approaches have been followed by researcher, enhancing the scalability. Here are the few strategies of application scaling in cloud.

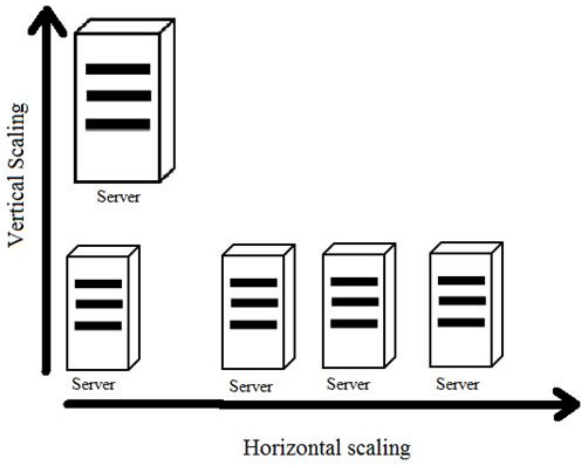

Figure 3 Scaling

Application scalability issues and strategies in different Service model of cloud is explained by Vaquero, Luis M., Luis Rodero-Merino, and Rajkumar Buyya [3]. Scaling in IaaS can be done in two ways- horizontal scaling and vertical scaling. Here author has classified two ways of achieving scalability in PaaS model- the container and the database management system (DBMS) scalability. User component is deployed and run in the software platform called container. Different platforms can be used by different PaaS providers. Author suggested that for the components hosted by PaaS, platform type can define different lifecycles, services and APIs. On the other hand, databases provide data persistence support. It should address the demand for data transactions support combined with big availability and scalability requirements.

The auto-scaling problem for a more general application model is explored and allow individual (non-uniform) job deadlines [4]. Here the aim is to complete all jobs by using resources that are less costly within the given deadline. Deadlines are used as a performance requirements specified by the users, and deadline misses are not strictly prohibited. Web request response time, network latency and a program's running time is selected as a deadline .Deadline assignment techniques are used to calculate an optimized resource plan for each job and the number of instances using the Load Vector idea is determined. Job scheduling and resource scaling is addressed at the same time by considering both the job-level and global-level cost-efficiency.

A lightweight approach is proposed by Han, Rui, et al for achieving more cost-efficient scaling of cloud resources at the IaaS cloud provider's side is proposed [7]. The multi-tiered applications, which are already implemented using multiple VM's have improved resource utilization between VM's as application demands vary. It proposed the following: Fine-grained scaling approach, Improving resource utilization, Implementation and experimental evaluation. They have proposed an intelligent platform based on the LS (light weight scaling) algorithm, which is implemented to automate the scaling process of cloud applications.

Automatic scaling problem is summarized in the cloud environment by Sharma, Tejinder, and Vijay Kumar Banga [18]. It is modelled as a modified Class Constrained Bin Packing (CCBP) problem. Here each server is a bin and each class represents an application. Here an innovative auto scaling algorithm to solve the problem and present a rigorous analysis on the quality of it with provable bounds is made. Compared to the existing Bin Packing solutions, item departure is supported which can effectively avoid the frequent placement changes caused by repacking. It support green computing by adjusting the placement of application instances adaptively and putting idle machines into the standby mode.

4. CONCLUSION

In this paper, we have surveyed various scaling strategies and load balancing techniques for cloud computing. Different techniques suggested by authors are discussed in this paper. The main purpose of scaling is adjust the application instances as per the user requirement. Scaling should take into consideration load balancing, by distributing load dynamically among the nodes and to make maximum resource utilization by reassigning the total load to individual node. Thus efficient and even distribution of resources is done, thus improving the performance of the system.

5. ACKNOWLEDGEMENT

I would like to express my gratitude to my project manager Arvind Jagtap for the useful comments, remarks and engagement through the learning process of this part of master thesis. Furthermore I would like to thank Bhaskar Kulkarni for introducing me to the topic as well for the support on the way. This research is sponsored by SAS R&D India and the contents of this paper belong to SAS R&D India.

References

[1] Zhang, Qi, Lu Cheng, and Raouf Boutaba. "Cloud computing: state-of-the-art and research challenges.", Journal of internet services and applications 1.1 (2010): 7-18.

[2] NIST: http://www.nist.gov/itl/cloud/

[3] Vaquero, Luis M., Luis Rodero-Merino, and Rajkumar Buyya, "Dynamically scaling applications in the cloud", ACM SIGCOMM Computer Communication Review 41.1 (2011): 45-52.

[4] Mao, Ming, and Marty Humphrey,"Auto-scaling to minimize cost and meet application deadlines in cloud workflows", Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, 2011.

[5] Bala, Anju and Chana, Inderveer, "A survey of various workflow scheduling algorithms in cloud environment", 2nd National Conference on Information and Communication Technology (NCICT), 2011.

[6] Kepes, Ben. "Understanding the cloud computing stack: SaaS, PaaS, IaaS ", Diversity Limited (2011): 1-17.

[7] Han, Rui, et al. "Lightweight resource scaling for cloud applications", Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on. IEEE, 2012.

[8] Chaudhari, Anand and Kapadia, Anushka, "Load Balancing Algorithm for Azure Virtualization with Specialized VM", International Journal of Innovations in Engineering and Technology (IJIET) Vol. 2 Issue 3 June 2013.

[9] Nayandeep Sran,Navdeep Kaur , "Comparative Analysis of Existing Load Balancing Techniques in Cloud Computing ",vol 2,jan 2013.

[10] Sumit Khurana, Anmol Gaurav Verma,"Comparison of Cloud Computing Service Models: SaaS, PaaS, IaaS",IJECT Vol. 4, Issue Spl - 3, April - June 2013.

[11] Rajan, Rajesh George, and V. Jeyakrishnan. "A Survey on Load Balancing in Cloud Computing Environments", International Journal of Advanced Research in Computer and Communication Engineering 2.12 (2013).

[12] Desai, Tushar, and Jignesh Prajapati. "A Survey Of Various Load Balancing Techniques And Challenges In Cloud Computing", International Journal of Scientific & Technology Research 2.11 (2013).

[13] Priya, S. Mohana, and B. Subramani. "A New Approach For Load Balancing In Cloud Computing", International Journal Of Engineering And Computer Science (IJECS-2013) Vol 2 (2013): 1636-1640.

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

[15] Ghuge, Kalyani, and Minaxi Doorwar. "A Survey of Various Load Balancing Techniques and Enhanced Load Balancing Approach in Cloud Computing", International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 10, October 2014.

[16] Hareesh M J , John P Martin , Yedhu Sastri , Anish Babu S ,"A Review on Load Balancing Algorithms in Cloud", IJCTA | March-April 2014.

[17] Shagufta Khan,Niresh Sharma ,"Effective Scheduling Algorithm for Load balancing (SALB) using Ant Colony Optimization in Cloud Computing", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)Volume 4, Issue 2, February 2014.

[18] Zhen Xiao, Senior Member, IEEE, Qi Chen, and Haipeng Luo, "Automatic Scaling of Internet Applications for Cloud Computing Services", IEEE Transactions On Computers, Vol. 63, No. 5, MAY 2014.

[19] Harshit Gupta, Kalicharan Sahu ,"Honey Bee Behavior Based Load Balancing of Tasks in Cloud Computing", International Journal of Science and Research (IJSR), Volume 3 Issue 6, June 2014.

[20] Joshi, N. A. "Dynamic Load Balancing In Cloud Computing Environments", International Journal of Advanced Research in Engineering and Technology (IJARET), Volume 5, Issue 10, October (2014), pp. 201-205