Volume 6, Issue 2, February 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review on Various Existing Load Balancing Approach in Cloud Computing

Shivani Goel Tripti Arjariya

Abstract: ***Cloud is the one of the fastest growing technology in era of computer science. It provides on-demand services (computing resources) to the users and demand for these computing resources is increasing rapidly. To fulfill this user demands virtualization is use. Virtualization is the core technology in the cloud which allows the sharing of the physical resources. It's enabled the service provider to create multiple virtual machines in a single physical machine. One of the important features of the virtualization is the VM migration, which allows transferring the VM from one PM to another PM. Hence, VM is the fundamental unit of the cloud which runs user application. Load on the VM is change constantly according to the application requirement. Due to this there may be a situation where some PM is overloaded whereas some VM are underloaded which degrades the performance of the PM. Energy consumption is one of most challenging issue in the cloud which is depends on the number of running servers. After reviewing the theory of cloud it is found that number of active server is depend on the resource utilization. In order to increase the resource utilization and reduce the power consumption, data centers needs an effective load balancing approach which distributes the load properly. Hug amount of work has been done in the field of load balancing. This paper explains various existing load balancing approach with their anomalies.***

Keywords: ***Distributed Computing, On Demand Resources, Cloud Computing, Virtualization, Server Consolidation, Load Balancing.***

I. INTRODUCTION

Cloud is the one of the fastest growing technology in era of computer science [1]. It's become to famous because of their attractive service such as easy to use, flexibility, cheapest etc. According to the NIST definition [2] "Cloud computing is 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" . It supports for the three types of services and can be deploy in four different ways [3, 4].

As show in figure 1 cloud can be deploy in four different way i.e. private, public, hybrid and community. Private cloud is a cloud which is use by single organization with in the network. User can not access the cloud services from outside the network. Whereas public cloud access anywhere in the world.

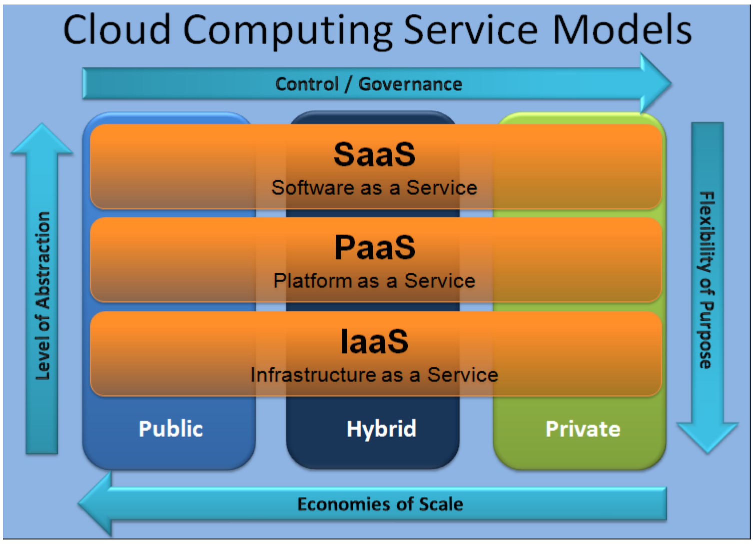

Figure 1: Cloud Computing Model

Community cloud is a cloud which is share by the multiple organizations and hybrid cloud is a combination of one or more cloud. Cloud supports three types of services Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). In SaaS, only software or applications are provide to the client as a service. Client use these software without ant installation such as Gmail, Facebook. But user does not control the operating system, hardware or network infrastructure on which it's running. In PaaS all computing resources such as hardware, software, network etc. are provide as a service. It is mainly use by the developer not by the normal user. In this type of cloud user does not have any control on the operating system, hardware or network infrastructure on which they are running. Whereas in IaaS all computing resources such as hardware, software, network etc. are provide as a service and user have full control on the operating system, hardware or network infrastructure on which they are running. These services are made available to the users through virtualization techniques. Virtualization [5, 6] is the backbone of the cloud computing. It is technologies which divide the physical resources and allow creating multiple VM in a single PM. Virtualization is implemented through the hypervisor, which is installed in each physical server and responsible for monitoring the resource utilization.

There is a great demand of virtualization in dynamic resource management. It reduces the server sprawl, minimize power consumption, balances load across physical machines. Virtualization technology is so flexible that it transfers the data when a machine hardware is overloaded, thus making easy work for hardware. In cloud when user request for the resources, virtual machine manager has to decide which host is suitable for hosting this new VM? This decision is taken on the basis of requirement constraint such as energy consumption, service level agreement (SLA) violation, downtime etc. VM placement is a NP hard problem [6]. Proper placement of the VM is a very important task in cloud, because wrong PM selection may increase the number of active server which lead to increase the energy consumption and increase the number of migration. Due to the dynamic nature of the VM, this PM selection is very challenging task. This paper explains some exiting VM placement approach with their anomalies.

II. LECTURE REVIEW

G. Xu et al. [7], proposed a load balancing approach for the public cloud based on the cloud partition. This approach first divide the entire public cloud into the partition and then balance the PM. They use the main controller and all other partition are to this main controller. When any users make request for the services, this request is reach to the main controller which decide which partition receive the job. The partition load balancer then decides how to assign the jobs to the nodes. This method minimizes the total migration time, but may increase the number of active server due to cloud partition.

Mayank Mishra and Anirudha Sahoo [8], proposed vector based load balancing approach for the cloud. In this approach they use various vector to represent the physical and virtual machine. All resource related information of the physical and virtual machine are represents in the form of vector. Total capacity of the PM is express by the Total Capacity Vector (TCV). Current utilization of the PM resources is represent by the Resource Utilization Vector (RUV). Remaining Capacity is represents by the Remaining Capacity Vector (RCV), which is the vector difference between TCV and RUV. Figure 2 shows the various vector used.

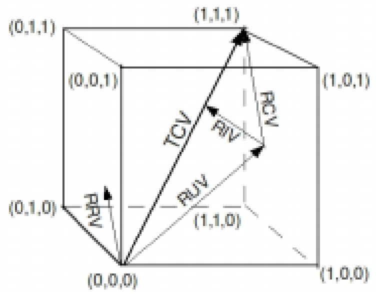

Figure 2: Depiction of Various Vectors used

A PM is balanced if RUV of a PM exactly aligns with the TCV. In this approach they place the VM in such a way that difference between the RUV and TCV is minimum. This approach may minimize the resource wastage, but did not focus on the energy consumption and only theoretical view of the approach is given.

R. Addawiyah et al. [9], proposed a load balancing approach using VM migration. To balance the PM or system lower and upper threshold are use. When the utilization of the PM reaches above the upper threshold PM is overloaded similarly if the PM utilization is below the lower threshold then PM is under loaded. In both cases VM migration is use. They set the value of lower and upper threshold are 10 and 90 respectively. When the PM is overloaded higher utilize VM is select for the migration and similarly when the PM is under utilize all VM running on that PM are chosen for the migration. These select VM is place to the PM where it consumed less power. Main objective of this approach is to minimize the energy consumption. For this purpose they place the selected VM to the PM where the energy consumption is minimum. This approach minimized the energy consumption but increase the total migration time.

K. Dasgupta et al.[10], proposed Genetic algorithm for the load balancing in cloud. Fitness function is use to assign VM to the PM. This method selects the set of VM randomly and then calculates the fitness function for each VM. VM with highest fitness value is assign to physical machine. To represent the current status of processing unit utilization processing unit vector(PUV) is use which is define as

$$PUV = f(MIPS, \alpha, L)$$

Where

MIPS is the million instructions per second

αis cost of execution instruction

L is delay cost

Job submit by the user is also represent by the vector called job unit vector (JUV), which is define as:

$$JUV = f(t, NIC, AT, wc)$$

Where

t represents type of service required by the job

NIC represents number of instructions present in the job

Job Arrival Time (AT)

Worst case completion time(wc) is the minimum time required to complete the job by a processing unit

After defining the PUV and JUV next they calculate the fitness value by using the following equation

$$\zeta = w1 * \alpha(NIC \div MIPS) + w2 * L$$

Where

w1and w2 are the weighting coefficient.

This method is easy to use and experiment result shows that it gives minimum response time as compare to other approaches i.e., first come first serve (FCFS), round robin etc. But this approach is suffer by the starvation. VM with higher fitness value is assign to the PM, so there is a possibility where VM with lower fitness value is in starvation.

Mayur S. Pilavare et al. [11], proposed load balancing approach based on the Genetic algorithm in cloud computing. Main problem with the genetic algorithm is the starvation. To mitigate this issue they assign priority to the VM and then use genetic algorithm for the VM placement. To assign the priority to the VM Logarithmic Least Square Matrix Technique [12] is use, which first calculat the comparison metrics for the each VM and then takes multiplication of all values in row then takes nth root of that product for all rows and normalizes the values. After this assign the highest priority VM to PM.

Following algorithm is us to calculate the priority to the VM.

Step 1: Randomly decide total number of processors.

Step 2: Calculate the comparison matrices for all processors.

Step 3: calculate the priorities of all processors according to the calculations of logarithmic square method.

Step 4: allocate the processor with highest priority with jobs to execute.

Step 5: End.

This method avoids the starvation issue. This method mainly focuses to minimize the response time. Since cloud is the dynamic in nature, so there are several issue that also must be focused during the design of any load balancing approach i.e., resource balancing, VM migration etc.

III. CONCLUSION

Cloud is a business model which provides the various services to the user. Due to its attractive features it's become so popular and size of the cloud is increasing very rapidly. Since cloud is a dynamic in nature where user requirement for the resources change frequently, so proper utilization of the resources is became a challenging task. For the proper utilization of the resources various load balancing approaches have been proposed by the various researchers. After reviewing the lecture of the cloud it is conclude that load balancing in cloud is a challenging task. This paper shows various load balancing approach in cloud with their anomalies.

REFERENCES

[1] R. Buyya et al., “Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility", Future Generation Computer Systems, 2011.

[2] Mell, P. et al.”The NIST Definition of Cloud Computing”, NIST Special Publication, 2011.

[3] R. K. Gupta et al., “A Complete Theoretical Review on Virtual Machine Migration in Cloud Environment”, International Journal of Cloud Computing and Services Science (IJ-CLOSER), Vol.3, No.3, June 2014, pp. 172-178.

[4] Sosinsky et al.” Cloud Computing Bible”, Wiley Publishing Inc 2012.

[5] C. Clark et al., “Live migration of virtual machines,” in Proceeding NS DI’05 Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation, vol. 2, pp. 273– 286, 2010.

[6] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, “Xen and the art of virtualization,” in Proceedings of the nineteenth ACM symposium on Operating systems principles, ACM, pp. 164 –177, 2003.

[7] G. Xu et al., "A Load Balancing Model Based on Cloud Partitioning for the Public Cloud", Journal of Tsinghua Science and Technology (IEEE), Volume 18, Issue 1, pp. 34-39, 2013.

[8] M. Mishra and A. Sahoo, "On Theory of VM Placement: Anomalies in Existing Methodologies and Their Mitigation Using a Novel Vector Based Approach", IEEE 4th International Conference on Cloud Computing, July 2011, pp. 275-282.

[9] R. Addawiyah et al., "Virtual Machine Migration Implementation in Load Balancing for Cloud Computing", six IEEE international conference, 2014

[10] K. Dasgupta et al. " A Genetic Algorithm (GA) based Load Balancing Strategy for Cloud Computing", First International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) (Elsevier), pp. 337-340, 2013

[11] Mayur S. Pilavare and Amish Desai, " A Novel Approach Towards Improving Performance of Load Balancing Using Genetic Algorithm in Cloud Computing", IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS'15, pp. 1-4 , March 2015.

[12] Adamcsek, Robert Fuller, A Thesis work on "The Analytic Hierarchy Process and its Generalizations" Eotvos Lorand University 2008.