A Review on Different Existing Load Balancing Approaches for Cloud

Saurabh Jain[1], Dr. Varsha Sharma[2]

[1]*Student, School of Information Technology, RGPV, Bhopal, MP, India*
[2]*Assistant Professor, School of Information Technology, RGPV, Bhopal, MP, India*

---***---

Abstract - *Cloud computing is one of the emerging technologies in modern IT era. It provides on demand IT related resources to the client on the rent basis. The key concept behind cloud computing is "virtualization". Virtualization is an interface which separate hardware from software and provides the benefits of server consolidation and live migration. A main benefit of virtualization is that, it allows creating multiple virtual machines (VMs) on a single physical machine (PM). Since multiple users share the cloud resources, so there may be situation when the PM is overloaded or under loaded. VM migration is a unique solution to mitigate both situations. Performance of the PM degrades when the PM is overloaded or under loaded. If load balancing in cloud is not handled properly, then it will increase the energy consumption and minimize the provider profit. Hence, load balancing in cloud is the prime requirement of the cloud provider. Load balancing in cloud is a challenging task due to change in user requirement at the run time. When the PM is overloaded or underloaded, some VMs need to be migrated. VM migration consists of three steps: source PM selection, VM selection and the last step target PM selection. Numerous load balancing algorithms have been proposed in the literature. This paper discussed various existing load balancing approaches with their comparisons.*

Key Words: **Virtualization, migration, energy efficient, virtual machine, physical machine**

1. INTRODUCTION

Cloud computing is one of the fascinating technology in the field of computer science. It became so popular due to its attractive features like on-demand services, utility based model, ease of use etc. According to Gartner's report [1], the Core technology behind the cloud computing is the virtualization [2, 3]. Virtualization is the abstraction layer between the hardware and software. It divides the physical resources into the multiple types and allows the sharing of physical resources. Virtualization increases the resource utilization because multiple users share the same physical resources but, it also introduces the need of load balancing. If the load is distributed properly then it will increase the resource utilization and reduce the energy consumption. During the study of load balancing approaches it is found that if the load is distributed effectively then it will minimize the energy consumption and total simulation time. Load balancing approach can be static or dynamic. Static approach is more appropriate for the cloud. The one where information about system is not important and working is less complex comes under static scheme whereas in the dynamic approach load balancing decision is taken based on the previous system information.

Most of the existing load balancing [4-6] approaches use thresholds to represent the overloaded and underutilized situation where lower threshold represents the under loaded situation and upper threshold is use to define the overloaded situation. To handle the overloaded and under loaded situations VM migration is used which move the VM from one PM to another PM. VM migration is the important feature of virtualization and is used to deal with the overloaded, under loaded and hot spot situations. Hence, VM migration allows the flexibility to resource provisioning. Several load balancing approaches have been proposed in the last few years. This paper gives the overview of some approaches with their anomalies.

2. Following are the Goals of load balancing in Cloud computing

1) To improve the performance of the cloud services.
2) To have a backup plan in case the system fails even partially.
3) Increase the availability of the resources.
4) To maintain the system stability.
5) To minimize the number of SLA violations.

3. Challenging Issue in load balancing

Main challenging issues for the load balancing are [3]

1) The load balancing approach imposed some overhead to process the request because it needs some resources for the processing. So this overhead should be minimize as possible
2) Since load in the cloud change dynamically, so it is very difficult to design an efficient load balancing approach for the cloud environment.
3) To balance the system, scheduler migrates the VM which reduce the system performance. So a load balancing approach must reduce the number of migrations.

4) Load balancing algorithms must be designed in the simplest possible forms, this is one of the challenging task for the developer.
5) Efficiency of the load balancing approach depends on the lower and upper threshold value. So selecting the proper value of the threshold is very difficult.

4. Related Work

R. Addawiyah et al, proposed a load balancing approach based on the threshold. They used two thresholds named lower and upper threshold for scheduling and migrating the VM. If the current resource's CPU usage is greater than 90% (which means that it is overloaded), then the virtual machine will be migrated to another resource with the CPU usage that is less than 50%. If the current resource's CPU usage is less than 10% (which means that it is underutilized), then the virtual machine will be migrated to another resource with CPU usage that is less than 70%. After selecting the overloaded or under loaded host, they select largest VM for the migrations. Main limitation of this approach is that it will increase total migration time as it selects the higher utilize VM for the migration [7].

X. Gaochao et al, present a load balancing approach which places the VM according to the partition. This approach first divide the data center in to the partitions according to the distance and then assign the VM according to the near partition. In each partition PM is divided into three categories named idle, normal and overloaded. To find in which categories host belongs load_ degree is used which is given by following equation

$$\text{load\_degree}(N) = \sum_{i=1}^{m} \propto_i * F_i$$

$$load\_degree_{avg} = \sum_{i=1}^{n} \frac{load\_degree(N_i)}{n}$$

Where
$\propto_i$ is the waiting coefficient ($\sum_{i=1}^{n} \propto_i = 1$)
N represents the current PM.
m is the different type of resources

Based on the value of $load\_degree_{avg}$ status of the PM is determined. This approach may increase the number of active servers due to partitioning of the datacenters [8].

Y. Fang et al, proposed a layered based VM scheduling approach for the cloud. According to this approach, VM scheduling approach consists of three layers named application layer, platform layer and infrastructure layer. Figure 1 shows the layered architecture used by this approach [10].

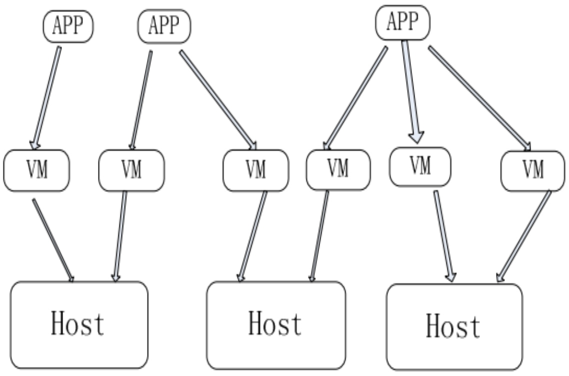

Figure 1: Two Level VM Scheduling Model

Application layer is designed for the users. Users can only interact with the application layer and submit task to the application layer. Result is also received by the user through this layer. Platform layer is a set of hardware and software which is used by the developer to design new application whereas platform layer is a set of virtual resources which is assigned to the user. First level scheduling between users and VM, finds the requirement of the VM and the second level scheduling between VM to host assign VM according to the VM description in the first level. This approach selects the VM on the first come first serve basis and assigns it to the PM which is lightest. Due to assigning the VM to the lightest PM this approach may increase the number of active servers [9].

M. Mishra et al, proposed a vector based VM scheduling approach for the cloud environment. All parameters which are required during the VM placement are represented by the vector. Resources required by the VM are represented by the RRV, resource utilization of the PM is represented by the RUV, and total capacity of the PM is express by the TCV etc. Figure 2 shows the various vectors used in this approach.

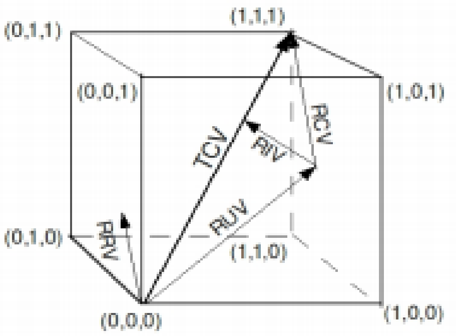

Figure 2: Depiction of Various Vectors used

This approach is mainly focused on the resource balancing and placing the VM to the PM where resource utilization vector of the PM is complementary to the resource requirement vector of the VM. Load on the PM depends on the number of VMs inside the PM which is calculated by the following equation:

$$HL_k = \sum_{i=1}^{n}(w_i) * \frac{\sum_{j=1}^{m}(RU_{ij})}{(Hcap_{ik})}$$

Where, HL_k is the load on kth host, n is the type of resources i.e. cpu, memory, ioetc, m is the number of VM in kthhost, RU_{ij} is the ithtype resources use by the jthVM and $Hcap_{ik}$is the ith resource capacity of kth host. This approach increases the resource utilization but only theoretical approach is given. That means practical implementation is not given in this approach [11].

Ekta Gupta et al. introduced an Ant Colony Optimization technique for the load balancing in cloud. This technique notice overloaded and under loaded servers and performs load balancing between identified servers of data center. This technique can achieve availability, effective resource utilization; cloud handled maximum number of requests and minimizes the response time. This approach is based on the training of the scheduler, so the efficiency of the load balancing approach entirely depends on the training which is a very challenging task due to dynamic nature of the user resources demands [12].

Mohammad H. AL et al. work on integrated algorithm that takes different server's load and effectively migrates the virtual machines. They used important factor for reducing energy cost as well as cooling cost at data centers by migrating virtual machines such that energy consumption by virtual machines is minimized and also switching off underutilized servers. When the load on the PM goes below the lower threshold then scheduler shutdown the PM to save the power consumption. But when the resource demand is increased then the PM is activated again which will take time. This situation can be avoided by switching the PM into the power saving mode rather than shut down [13].

Gaston Keller et al. introduced a problem in consolidation environments which is how to deal when number of virtual machines demand exceeds the resource capacity of the host. If we know better that which virtual machine to migrate and which hosts to migrate them, it can ease the situations. In his paper authors proposed a First Fit-based relocation policies, which reckon hosts and virtual machines in dissimilar order. Authors present the simulation results that exhibit the policies depending on the scenario and the metrics ascertained [14].

G. Shobana et al. proposed aim of load balancing as "to remove overload of any of the resources, maximize throughput, and minimize response time". According to authors for achieving these we need to do load balancing properly. This paper introduced the preemptive task scheduling algorithm that almost abbreviate make span which observe honeybee's foraging behavior. Aim of this algorithm is to maximize throughput and minimize latency by priority of the tasks [15].

Table -1: Comparison of Different Load Balancing Approach

Approach	Description	Advantage	Disadvantage
Virtual Machine Migration Implementation in Load Balancing for Cloud [7]	It is a threshold based load balancing approach which uses migration to balance the PM.	➢ Minimize the number of migration ➢ Easy to implement	➢ Increases total migration time ➢ Increases number of active server
A Load Balancing Model Based on Cloud Partitioning for the Public Cloud [8]	This approach place the VM according to the partition or location.	➢ Reduce the total migration time ➢ Reduce the total transmission time	➢ Increases number of active server
Design and Implementation of an efficient Two-level Scheduler for Cloud Computing Environment [9]	This approach present two layer model for the VM placement	➢ Increase physical and virtual machine performance	➢ Increases the energy consumption ➢ Threshold is not defined

On Theory of VM Placement: Anomalies in Existing Methodologies and Their Mitigation Using a Novel Vector Based Approach [11]	This approach uses vector method for place the VM.	➢ Reduce the resource wastage ➢ Increase the PM performance	➢ Only theoretical model for load balancing is given ➢ Threshold is not defined
A Technique Based on Ant Colony Optimization for Load Balancing in Cloud Data Center [12]	It is a load balancing solution using Ant Colony approach.	➢ Achieve availability ➢ Effective resource utilization ➢ Minimize response time	➢ Increases the energy consumption ➢ Increases number of migration
An Energy-aware Virtual Machine Migration Algorithm [13]	It is an energy aware load balancing approach.	➢ Easy to implement ➢ Minimize the energy consumption	➢ Increases the total simulation ➢ Degrades the PM performance
An Analysis of First Fit Heuristics for the Virtual Machine Relocation Problem [14]	This approach offer solution for which virtual machine to migrate and to which hosts to migrate them	➢ Minimize response time ➢ Easy to implement	➢ Uses first fit which increase the number of running servers ➢ Increases the energy consumption

5. Conclusion

Load balancing in cloud is a challenging task due to the change in user requests at run time. After reviewing the theory of load balancing in cloud it is found that VM migration process is used for load balancing of PMs. Lower and upper thresholds are used to define the overloaded and under loaded situation of the PMs. When the load on the server is less than the value of lower threshold, system is said to be under loaded. Similarly when load on the server is more than the value of upper threshold, system is said to be overloaded. In both cases some VMs have to migrated to balance the system load. Static threshold is more suitable for the cloud as compare the dynamic threshold because the load on the PM is changed very frequently. This paper discussed some existing load balancing approaches with their comparison.

REFERENCES

[1] R. Hunter, The why of cloud, http://www..gartner.com/DisplayDocument?doc_cd=226469&ref= g_ noreg, 2012.

[2] M. D. Dikaiakos, G. Pallis, D. Katsa, P. Mehra and A. Vakali, “Cloud Computing: Distributed Internet Computing for IT and Scientific Research”, IEEE Journal of Internet Computing, Vol. 13, No. 5, 2009, pages 10-13.

[3] B. P. Rima, E. Choi and I. Lumb, “A Taxonomy and Survey of Cloud Computing Systems”, Proceedings of 5th IEEE International Joint Conference on INC, IMS and IDC, 2009, pages 44-51.

[4] D.Perez and Botero, “A brief tutorial on live migration of virtual machine from a security perspective”, 2011

[5] C. Clark, K. Fraser, S. Hand and J. C. Warfield,“Live migration of virtual machine”, proceeding of the 2nd international conference on symposium on network system design and implementation, 2007 pp. 273-286.

[6] GunjanKhanna, Kirk Beaty, GautamKar, and AndrzejKochut, “Application performance management in virtualized server environments”, proceeding of the 10th IEEE conference on Network Operations and Management Symposium, 2006, pp. 373-381.

[7] Rabiatul Addawiyah, Mat Razali, Ruhani Ab Rahman, Norliza Zaini and Mustaffa Samad, "Virtual Machine Migration Implementation in Load Balancing for Cloud Computing", 5th International IEEE Conference on Intelligent and Advanced Systems, 2014, pp. 1-4.

[8] GaochaoXu, Junjie Pang, and Xiaodong Fu, “A Load Balancing Model Based on Cloud Partitioning for the Public Cloud” in the proceeding of IEEE conference on Tsinghua Science and Technology, 2013, pp. 34-39.

[9] Yiqiu Fang, Fei Wang and JunweiGe, "A Task Scheduling Algorithm Based on Load Balancing in Cloud Computing", 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, 2010, pp. 271-277.

[10] Sadhasivam Dr., S., Jayarani, R.: Design and Implementation of an efficient Two-level Scheduler for Cloud Computing Environment. In: International Conference on Advances in Recent Technologies in Communication and Computing, vol. 148, pp. 884–886 (2009).

[11] M. Mishraand A. Sahoo, "On Theory of VM Placement: Anomalies in Existing Methodologies and Their Mitigation Using a Novel Vector Based Approach", 4th IEEE international conference on cloud computing in 2011, pp. 275 - 282.

[12] Ekta Gupta and VidyaDeshpande, "A Technique Based on Ant Colony Optimization for Load Balancing in Cloud Data Center", proceeding of the 13th International Conference on Information Technology, December 2014, pp. 12-17.

[13] Mohammad H. ALShayejiand M.D. Samrajesh, "An Energy-aware Virtual Machine Migration Algorithm", proceeding of the International Conference on Advances in Computing and Communications, August 2012, pp. 242-246.

[14] Gaston Keller, Michael Tighe, HananLutfiyya and Michael Bauer, "An Analysis of First Fit Heuristics for the Virtual Machine Relocation Problem.", proceeding of the 6th International DMTF workshop on Systems and Virtualization Management (SVM)/CNSM,October 2012, pp. 406 - 413.

[15] G. Shobana, "Nature Inspired Preemptive Task Scheduling for Load Balancing in Cloud Datacenter" proceeding of the IEEE International Conference on Information Communication and Embedded Systems (ICICES), Feb. 2014, pp. 1-6.