Virtual Machine Load Balancing: State-of-the-Art and its Research Challenges

Amit Yadav
M.Tech Scholar
B.T.K.I.T Dwarahat, Almora, Uttarakhand

Lalit Mohan Joshi
M.Tech Scholar
B.T.K.I.T Dwarahat, Almora, Uttarakhand

Archana Verma
Assistant Professor
B.T.K.I.T Dwarahat, Almora, Uttarakhand

ABSTRACT

Cloud computing is the most demanding technologies in today's era due to the reduction of cost and complexity of applications, as a result number of users demanding more services of cloud. Thus, for overall cloud performance load balancing in the cloud is an important research area. Cloud handles large number of users as it is working on powerful datacenters. So for better efficiency it has to be featured with reliable load balancer for handling the load in the cloud. Many load balancing algorithms have been proposed for the better management of user request among cloud resources. This paper presents our research with a survey of various existing load balancing algorithms along with state of the art and its research challenges.

Keywords- Cloud Computing, Load Balancing, Load Balancer, Static/Dynamic Load Balancing

1. INTRODUCTION

Cloud Computing is a network oriented environment for sharing resources and computation. In the last five years cloud came into popularity. It provides a flexible and easy way to use different resources over the network as a part of its service. The Cloud provides a better and flexible way to keep and retrieve large files and datasets for the users spreader across the world. Cloud categorized into three service models, i.e. Software as a service (SaaS), platform as a service (PaaS) and Infrastructure as a service (IaaS). In IaaS, all the physical resources are working for cloud environments. Cloud is basically known for its virtualization environment at IaaS level. Here physical resources when split into some logical units it is then called a virtual machine (VM). Some of the enterprise virtual machine frameworks are Xen [1], Kernel virtual machine (KVM) [2], VMware [3], Virtualbox [4]. A virtual machine in a cloud environment is operated and managed by some well known virtual machine monitor or Hypervisor [5]. Some common used hypervisors in cloud environments are Open stack [5], Eucalyptus [5] and Cloudstack [5]. These hypervisor manages all the resources at the IaaS level in the cloud. One of the main functions of the hypervisor is load balancing among virtual machines on a pool of cloud network. Here Load balancing in virtual machines is the process of allocating virtual machines to physical services according to the user need. Load balancing methods are designed to determine which VM is assigned for the next execution of task units. In Cloud resources are well organized for the efficient conductance of network services. So resource management plays a significant role in the efficient performance of the cloud environment. According to SLA(service level agreement)[6], load balancing algorithms in datacenter ought to make proper management of resources with the minimum service violation.

Load balancing can affect the cloud resources, performance which rely on the amount of work assigned to resources for a specific time period. So there is a need of load balancing among the resources**.** This paper is designed to provide better awareness of load balancing algorithm in virtual machines with research gap and challenges. We surveyed various proposed algorithm of virtual machine load balancing. The research challenges are then explained along with our future work. This paper is categorized as follows Why VM load balancing in Section 2.Survey of existing work in Section 4. Research challenges in load balancing in cloud in Section 5. We concluded the research work and future work in Section 6. Aneka Cloud is highlighted in Section 7.

2. WHY LOAD BALANCING IN CLOUD COMPUTING

Load Balancing is defined as a technique of distributing many computations and communication among various resources working on the network. So load balancing works on the proper satisfaction of the users by the proper resource utilization. So an efficient load balancing technique will be the one which will minimize the resource consumption [10]. In a cloud computing hypervisor act as a virtual machine monitor which works for the proper resource management as (Virtual machine sharing) in a cloud network. So various load balancing algorithm has been proposed for cloud environment and all the algorithm works for the improvement of the parameters such as

1. **Throughput**: - Sum of the complete execution of task is called throughput. SO if a throughput is high system will perform better**.**
2. **Related Overhead: -** During the execution of load balancing algorithm the amount of overhead produced. So a successful implementation of the algorithm means minimum overhead.
3. **Fault Tolerance: -** At any arbitrary node working in the system. It is the way to perform uniformly even at any fault at any arbitrary node in the network system.
4. **Migration Time: -** Time taken by the tasks in transferring from one machine to another machine in the system. Migration time should be minimum for better performance.

5. **Response Time: -** Minimum time that a Load balancing algorithm takes to respond in the distributed system.
6. **Scalability:-** It depicts the capability of the system to complete the load balancing algorithm having less consumption of resource(processor or machine).

Green Computing in the cloud can be achieved by Load balancing by reducing energy consumption from network resources.[7]

2.1 Classification of Load Balancing Algorithm

Load Balancing algorithm classified into two forms on the basis of the state of the system.

2.2 Static Algorithm

The static algorithm works on the prior information about the resources of the system and applications so that it can allocate the load to virtual machines in an efficient manner. In this each resource has to be static in nature. The static algorithm divides the network traffic equally between VM's. These algorithms are non-preemptive in nature. Round Robin [8] is under this category, but due to some critical issues associated with round robin, weighted round robin [8] is introduced. CloudSim [9] has introduced weighted round robin in a simulative manner. Each VM has been assigned static weight and VM has higher weight will take more connection [10]. Central load balancing decision model [11] suggested by Radojeyik is also an enhancement of round robin.

2.3 Dynamic Algorithm

Dynamic load balancing works on the dynamic changes in the state of the system in terms of nodes capabilities and network bandwidth. It focus on the procedure to switch the processor from an over utilized machine to underutilized so that execution will be fast, It works on the preemptive scheduling for its dynamic nature. For this nature algorithm should require constant node monitoring and should maintain task updating which is usually harder to implement. Many algorithms have been proposed under the INS (Index name server) for the integration of duplication of the task and efficient section of access node.

3. SURVEY OF EXISTING WORK

We have surveyed various load balancing algorithm techniques which are currently being worked on various cloud computing environments and on the basis of performance analysis. This will be beneficial for the researchers carry their further research work in the relevant area.

Compare and Balance

Platform used- Infrastructure Cloud

It works in two phases. One is to sample the entire infrastructure of virtual machines. Second is to use the adaptive runtime migration of virtual machines. It is efficient as it balances the load to equilibrium very fast. It migrates the virtual machine from high cost physical servers to low cost[12].

Honeybee foraging Behavior

Platform used - Large scale cloud system

It is a distributed load balancing algorithm which works dynamically to achieve global load balancing through considering local server. Its performance is high[13].

Biased-Random Sampling

Platform used - large scale cloud system

It does the random sampling of the jobs and by a local relocation of similar services having same job assignments[13].

Carton

Platform Used- Framework of cloud controller

It works on cloud controller [] to minimize the load balancing cost and does four assignments of resources by limiting the distributed rate. It is easy to deploy and has very low computational complexity[14]

LBVS

Platform used- Cloud Storage

A load balancing strategy for virtual storage. This algorithm is proposed for cloud storage. The exponential increase in data required higher data storage performance. LBVS uses fair-share replication method which maps the physical resources to uniform virtual resources which then assembles that virtual space into global ones[16]

Server based LB

Platform used- Distributed web server

To avoid remote server overloading this algorithm uses a protocol to limit the overloading rates and redirect it to another less loaded machine. It uses a heuristic approach to manage the load changes[16]

Join-Idle-Queue

Platform used-Cloud data center

This algorithm uses a queue based memory in which information about the idle processor is kept and if an availability request comes idle process is allotted and it reduces queue length[17]

4. RESEARCH CHALLENGES IN LOAD BALANCING

To extend our research and to propose an optimum solution for the issue of load balancing in the cloud we need to find out the main parameters of load balancing challenges which will help for a better proposed algorithm for our research work.

4.1 Spatial Distribution of Cloud Nodes

Communication delay is very less in some intranet network and many of the algorithm are deployed and tested for this platform only. However to work on special environment for a load balancing algorithm is a challenge because factors like congestion will increase between node to node connectivity. To tolerate the high delay [18] over network efficient load balancing algorithm has to be developed.

4.2 Data Storage Replication

A replication algorithm for load balancing is categorized as full and partial replication. Full replication provides some data in all nodes which will increase the cost of storage resources. So partial type will be better as it will save parts of the datasets in each node which depends on the processing capability [18]. This will enhance the better utilization yet it will increase algorithm complexity due to its partial nature. In VMware, Vsphere [4] is a program which uses a partial algorithm of replication.

4.3 Algorithm Complexity

Load Balancing has to be designed with less complexity in terms of implementation and operation. If the complexity will be higher in implementation it will create a more complex performance with some related issues.

4.4 Point of Failure

Some load balancing algorithm like centralized algorithms provides an efficient mechanism for solving several issues of load balancing. Here one controller is managing the load of the entire system. In such a case if the controller fails, the whole system will be affected. So every load balancing algorithm has to be designed to overcome from this problem [19]. The distributed dynamic algorithm is more complex to implement still it seems to provide the best approach.

5. FUTURE WORK

In this paper, we have surveyed various load balancing algorithms for cloud computing environment. We have presented the research challenges of virtual machine load balancing algorithms. On the basis of survey we concluded that we have to design an efficient algorithm that can tolerate high delay capable of different resources and equally distributed the task. So we will propose map-reducing technique which will deploy on the Aneka Cloud Platform.

Aneka Platform

Aneka[www.manjrasoft] is a cloud application platform which is a research project of Melbourne University developed under Manjrasoft Technologies. It supports various developments of programming models of scalable applications to run in many types of clouds in an elastic manner. It supports the programming abstraction for developing a runtime environment that can be deployed on various hardware (clusters, network and cloud resources) for the load balancing environment. The developer can choose different programming abstractions to design their application, task, distributed threads and map reduce. Their applications are then executed in the distributed service oriented runtime environment which can dynamically integrate additional resource on demand. These technologies are key examples of cloud computing.

6. REFERENCES

[1] Xen, (2015), [online]. Available: http://www.xenproject.org, [March, 21, 2015]

[2] Linux,(2014), “KVM 4.2 “, [online]. Available: http://www.linux-kvm.org, [Oct, 15, 2014]

[3] Sun-Oracle,(2015), “ Virtual Box 8.2”, [online]. Available: http://www.virtualbox.org, [March, 16, 2015]

[4] Vmware(2015), [online]. Available:http://www.vmware.org,[March,17,2015]

[5] National Institute of Standards and Technology,(2014), [online] Available: http://en.wikipdeia.org/wiki, [March,20,2015]

[6] L.Deboosere, V.Bert, S.Pieter, D.T.Filip,D.Bart and D.Piet, “Efficientresource management for virtual desktop cloud computing,” Springer, 2012.

[7] J.Hu,J.Gu, G.Sun, T.Zhao, “A Scheduling Strategy on Load Balancing of Virtual Machine Resources in Cloud Computing Environment,” Third International Symposium on Parallel Architectures,Algorithms and Programming(PAAP), pp. 89-96, 2010.

[8] Rimal, B. Prasad, E. Choi and I. Lumb, "A taxonomy and survey of cloud computing systems." In proc. 5th International Joint Conference on INC, IMS and IDC, IEEE, 2009.

[9] CLoudSim, (2015), [online]. Available: http://www.cloudbus.org, [March, 21, 2015]

[10] Gunarathne, T., T-L. Wu, J. Qiu and G. Fox, "MapReduce in the Clouds for Science," in proc. 2nd International Conference on Cloud Computing Technology and Science (CloudCom), IEEE, pp:565-572, November/December 2010.

[11] Z.Zhang, X.Zhang, “A Load Balancing Mechanism Based on Ant Colony and Complex Network Theory in Open Cloud Computing Federation,” Proceedings of 2nd International Conference on Industrial Mechatronics and Automation (ICIMA),Wuhan, China, pp. 240-243, 2010.

[12] Y. Zhao, W.Huang, “Adaptive Distributed Load Balancing Algorithm based on Live Migration of Virtual Machines in Cloud,” Proceedingsof 5th IEEE International Joint Conference on INC, IMS and IDC,Seoul, Republic of Korea, pp. 170-175, 2009.

[13] M.Randles, D.Lamb, A.Taleb-Bendiab, “A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing,” Proceedings of 24th IEEE International Conference on Advanced Information Networking and Applications Workshops, Perth, Australia, pp. 551-556, 2010.

[14] R.Stanojevic, R.Shorten, “Load balancing vs.distributed rate limiting: a unifying framework for cloud control,” Proceedings of IEEE ICC,Dresden, Germany, pp. 1-6, 2009.

[15] H.Liu, S.Liu, X.Meng, C.Yang, Y.Zhang, “LBVS: A Load Balancing Strategy for Virtual Storage,” International Conference on Service Sciences (ICSS), IEEE, pp. 257-262, 2010.

[16] A.M.Nakai, E.Madeira, L.E.Buzato, “Load Balancing for Internet Distributed Services Using Limited Redirection Rates,” 5th IEEE,Latin- American Symposium on Dependable Computing (LADC), pp. 156-165, 2011.

[17] Y.Lua, Q.Xiea,G.Kliotb, A.Gellerb, J.R.Larusb,A.Greenber, “Join-Idle- Queue: A novel load balancing algorithm for dynamically scalable web services,” An international Journal on Performance evaluation, In Press, Accepted Manuscript, 2011.

[18] Y. Zhao, W.Huang, “Adaptive Distributed Load Balancing Algorithm based on Live Migration of Virtual Machines in Cloud,” Proceedings of 5th IEEE International Joint Conference on INC, IMS and IDC,Seoul, Republic of Korea, pp. 170-175, 2009.

[19] H.Liu, S.Liu, X.Meng, C.Yang, Y.Zhang, “LBVS: A Load Balancing Strategy for Virtual Storage,” International Conference on Service Sciences (ICSS), IEEE, pp. 257-262, 2010.