ISSN: 2321-7782 (Online)
Volume 2, Issue 12, December 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Migration of Virtual Machines: A Way to Load Balancing in Cloud Environment

Syam Sankar[1]
PG scholar,
Department of Computer Science,
College of Engineering Perumon,
Kerala, India

Devi Dath[2]
Assistant Professor,
Department Of Computer Science,
College Of Engineering Perumon,
Kerala, India

Abstract: A group of interconnected servers with the responsibility of delivering resources to the end users by following 'pay as you use 'model constitutes a cloud system. As the cloud system grows, work load handled at the servers (or hosts) increases drastically. There should have a proper load balancing methods so that work load generated by the clients can be distributed smoothly across all servers. Load balancing across all servers is maintained by virtual machine migration. Cloud providers have to ensure the fast execution of jobs by implementing the methods of load balancing and thereby achieving quick response of submitted tasks and maximal utilization of hosted resources. This paper makes a study on the concepts associated with load balancing scenarios. The objective of this paper is to present the load balancing techniques in a simpler way and to analyze the issues of load distribution.

Keywords: Virtual machine, migration, overload, cloud, load balancing

I. INTRODUCTION

On-demand delivery of computing resources over the internet implies cloud computing. Computing resources include memory, storage, CPU, bandwidth etc. Instead of installing all softwares and hardwares on your PC, you will get it as a service from a company (cloud provider) over the internet and you have to pay (or free) for the use of those resources. Amazon, Citrix, Google, Microsoft etc. are the major cloud providers. Cloud interface softwares like web browser allow a user to connect to the system of cloud. Gmail, Yahoo etc. are following the concepts of cloud computing. We can login to our mail account from our PC, but the actual storage of data is made at a remote server.

According to the publication titled "*The NIST Definition of Cloud Computing*" by NIST (US), a cloud system is having five essential characteristics: resource pooling, on-demand self service, rapid elasticity, broad network access, and measured service. The cloud computing resource models are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [1]. The major resources managed at the SaaS layer are subscription based (on-demand) business applications, web services, multimedia, PPM applications etc. The giant providers of these resources are Google Apps, Gmail, Salesforce.com etc. Paas layer provides webservers, development tools, databases etc. Microsoft Azure, Google App Engine, force.com are PaaS providers. Storage systems, physical servers, virtual machines and networks are managed by IaaS layer. Amazon, GoGrid, Rackspace etc. are IaaS providers. Cloud system can function mainly in three ways: public cloud, private and hybrid cloud. Virtualization is the main functional component of cloud computing. It allows us to create multiple virtual machines within a host. Hypervisor is the software used to create virtual machines within a host. Available physical resources are shared by the created virtual machines. We can consider virtual machines as the processing units in cloud environment. The jobs generated from client side are submitted to the virtual machines and the resource requirement of jobs should meet at the assigned virtual machines. Sometimes virtual machines get overloaded, then tasks must be removed from overloaded virtual machines to undeloaded virtual machines [2] to distribute the load equally across all virtual machines. This is a review paper and it simply explains the concepts of load balancing in cloud environment.

II. Load Balancing Concepts

Load balancing is the process of distributing load across all servers such that no server should get overwhelmed with jobs or tasks. A cloud datacenter holds a set of servers (or hosts). Whenever a client made a request to do a particular job, then the request is directed to a server with sufficient resources to do that job. The request in turn is handled by a suitable virtual machine running on it. As the requests raised by clients increases, the load on the server also increases. Work load of a server simply means the number of running tasks on that server and the load on the server varies dynamically.

Resources are allocated dynamically to the applications based on the requirements. Overload comes when the resource demands of virtual machines where applications are running go beyond the threshold value of resources utilization. Each resource can be assigned a threshold value, indicting the maximum percentage of utilization. If the utilization or simply usage of a resource in a server exceeds its threshold value, then, it indicates a situation of overload in that server. Suppose CPU threshold is set to be 92% in a server. Whenever the CPU usage (CPU utilization) in that server goes beyond 92%, we can say the server is overloaded. High resource demands in a server make a situation of overload.

When a server is overloaded, i.e., tasks take long waiting time to get processed due to shortage of free resources for execution, then *live migration of virtual machines* is the only solution. We can call an overloaded server as a hotspot [3]. Consider a simple load balancing scenario [1]:

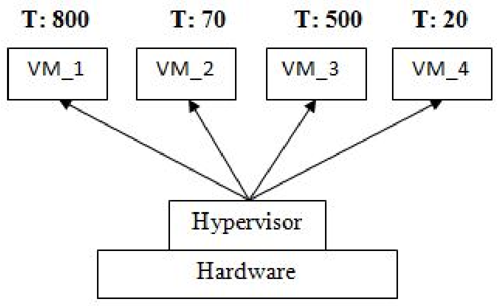

Fig.1 Server _1

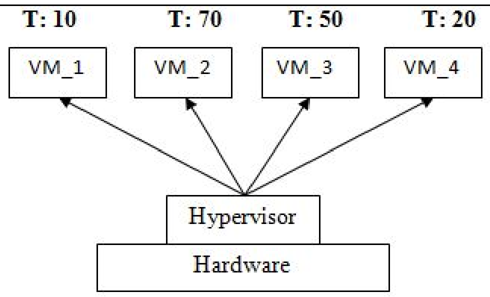

Fig.2 Server _2

Suppose we have two servers (Server_1 and Server_2) running in a cloud datacenter. Four virtual machines (VM_1, VM_2, VM_3, VM_4) running on each server. The letter '**T**' denotes the number of tasks (load) waiting in the queue at a virtual machine. Assume that Server_1 is overloaded since it is having so many tasks waiting to be processed and definitely the CPU utilization must be greater than its threshold. As it is clear from the figure that Server_2 is running smoothly by properly distributing load across all virtual machines and say it is undeloaded, that is all resource utilization lie beneath the level of its corresponding threshold. To reduce the load of Server_1, migration of virtual machines must be performed. A suitable virtual machine from Server_1 must be found so that it can be migrated to an underloaded server. From this situation, it is clear that VM_1 of Server_1 is the best one to be transferred from it since it is loaded with many number of tasks. After finding an overloaded virtual machine, here it is VM_1, migrate it to an underloaded server (Server_1).

Migration process [4] considers the following aspects:

1. A server can be labeled as *overloaded* if its resource utilization (CPU usage, memory usage, number of tasks etc.) exceeds its corresponding threshold values.
2. Consider all servers labeled with *overloaded*, make a sorted list on the virtual machines running on the servers based on load. The virtual machine (VM) that comes first in the list is the one which is loaded heavily with tasks. This list gives an order to select a virtual machine for migration.

3. Migrate the VM from an overloaded host to a suitable under loaded host using pre-copy or post-copy method
4. A destination server should not be overloaded after accepting a virtual machine from an overloaded host.
5. Ensures that the cost of migration is minimum

Migration of virtual machines causes the load on the server to get reduced.

III. MIGARTION OF VIRTUAL MACHINES

Virtualization softwares like Xen, VMware etc. allow to create multiple virual machines within a physical host. Each virual machine is designed with its own operating system and a set of applications running on it. All the virual machines share the underlying hardware and having virual resources like CPU, memory etc. provisioned by an abstraction layer created by the hypervisor. Encapsulation and hardware independence are the two important properties of a virtual machine. Due to these properties, a VM can be migrated easily from one host to another.

Migration of virtual machines should not be visible to an end user. If migration takes longer time due to some reason, then the user may feel service unavailability and that is unacceptable. Migration is done in such a way that it must reduce downtime. Process level migration faces so many difficulties. Now virtual machine migration is a good alternative to it since we migrate an entire OS and its applications as a single unit. Virtual machine migration also allows consistent memory (i.e. pages) transfer from one host to another. The stages of migration include [5] pre-Migration, Reservation, Iterative pre-copy, Stop-and-Copy, Commitment and Activation. The live migration [6] as opposed to *'stop and copy'*, which involves VM halting, copying memory pages from source host, and restarting new VM in destination, has the advantage to copy an entire VM unit (OS and its applications) with approximately zero downtime.

The cost of virtual machine live migration in cloud environment [6] is acceptable but cannot be disregarded, especially in systems where service availability and responsiveness are governed by strict Service Level Agreements (SLAs).

IV. CONCLUSION

Load balancing in a cloud environment is an inevitable procedure to be followed. Whenever a server gets overloaded, we can use the method of virtual machine migration to reduce the load. Load balancing attempts to improve the response time of the tasks submitted by the clients. Cloud computing is a very vast technology. Load balancing is the most challenging issue associated with it. There is a large scope to develop much more efficient algorithms in balancing the cloud load and thereby achieving speedy processing of client requests.

References

1. Syam Sankar, Devi Dath, Load Balancing in Cloud Environment Using Task Transfer: A Review (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014, 8127-8129 ISSN 0975-9646
2. Dhinesh Babu L.D, P. Venkata Krishna, Honey bee behavior inspired load balancing of tasks in cloud computing environments, Applied Soft Computing 13(2013)2292-2303,journal,homepage:www.elsevier.com/locate/asoc
3. Weijia Song, and Qi Chen, Zhen Xiao, Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 24, NO. 6, JUNE 2013
4. Varsha P. Patil and G.A. Patil, Migrating Process and Virtual Machine in the Cloud: Load Balancing and Security Perspectives Varsha P. Patil and G.A. Patil, International Journal of Advanced Computer Science and Information Technology 2012, Volume 1, Issue 1, pp. 11-19, Article ID Tech-21 ISSN 2320 – 0235
5. Ashima Agarwal, Shangruff Raina, Live Migration of Virtual Machines in Cloud, Ashima Agarwal, Shangruff Raina, International Journal of Scientific and Research Publications, Volume 2, Issue 6, June 2012 ISSN 2250-3153
6. W. Voorsluys, J. Broberg, S. Venugopal, and R. Buyya, Cost of Virtual Machine Live Migration in Clouds: A Performance Evaluation,W.Voorsluys, J. Broberg, S. Venugopal, and R. Buyya, Porc. 1st Inte'l. Conf. on Cloud Comput., Beijing, China, pp. 254-265, Dec. 1-4, 2009.