DYNAMIC LOAD BALANING BASED ON LIVE VIRTUAL MACHINE MIGRATION

Manh Do
manh.do226@topper.wku.edu
Department of Computer Science
Western Kentucky University

Dr. Michael Galloway
jeffrey.galloway@wku.edu
Department of Computer Science
Western Kentucky University

ABSTRACT

Recently, cloud computing is a new trend emerging in computer technology with a huge demand from the clients, which leads to the consumption of a tremendous amount of energy. Load balancing is taken into account as a vital part of managing income demand, improving the cloud system's performance and reducing the energy cost. Live virtual machine migration is a technique to perform the dynamic load balancing algorithm. To optimize the cloud cluster, there are three issues to consider: First, how does the cloud cluster distribute the virtual machine (VM) requests from clients to all physical machine (PM) when each machine has a different capacity. Second, what is the solution to make CPU's usage of all PMs to be nearly equal. Third, how to handle two extreme scenarios: rapidly rising CPU's usage of a PM due to sudden heavy workload requiring VM migration immediately and resources expansion to respond to heavy cloud cluster through VM requests. We also provide the implementation and results of this approach, which the performance of the cloud cluster is improved significantly.

BACKGROUND

Cloud
Cloud computing – the current generation of computing system. It's a model which combines grid computing and supercomputing. The main idea of cloud computing is a virtual service which can execute any type of job. The original of the term cloud computing is ambiguous, the word "cloud" in science is normally related to a humorous of object that is appear very far from client. Cloud computing concept appeared early as 1996 which was mention in a Compaq internal document. [1]

Load Balancing
Load balancing is the process of redistributing the entire workload among compute nodes of the cloud cluster in order to make resource utilization and reduce the response time. One of basic resource utilization concepts is to eliminate the scenario in which some of compute nodes have heavy workload while others are under-loaded or in idle. Actually, load balancing is done by VM migration or process migration.

Live Migration
In this type of migration, without turning off the VM, the running VM on the host machine is migrated to the destination host. There are two benefits of live migration are to improve uptime, reduce downtime and to go green by save energy.

Glusterfs:
Glusterfs is a scalable network filesystem suitable for data-intensive task such as cloud storage. In this research, all data are stored in a PM (storage node), others can access to this machine via Glusterfs. Sharing storage is a key to implement live migration.

DYNAMIC LOAD BALANCING

A. Distribute workload
Based on the request VMs from clients, head-node, a server which manages the whole cloud cluster will calculate itself to distribute the whole VM requests to all computer nodes depending on the capacity and the number running VMs in each computer node. There are few steps to be applied. First, head-node will distribute equally VM requests to each PM. Based on the capacity, each physical machine will try to execute as much as it can the requests. The number of lacking will be sent back to head-node and to be distributed to others in next iteration. After calculating exactly the number of VMs for each physical machine, controller of head-node will send instantaneous VM requests to all compute nodes at one time.

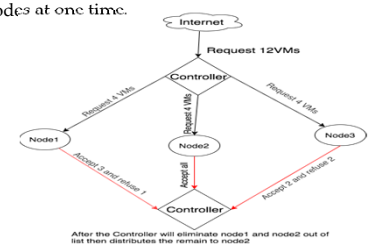

Fig 1: Distribute Worked Load

B. Optimize CPU's usage [3]
The cloud centers are mostly constituted from heterogeneously servers, which contain a different number of VMs due to fluctuating resource usage, may cause to imbalance resource degrade performance. To achieve efficient resource utilization, optimizing CPU's usage of the whole cloud data center is the most challenger for every load balancing algorithms. There are a lot of researchers who focus on this area but most of them do not apply the benefit of live VM migration. Dynamic load balancing takes more advances than the other. The goal of this algorithm is to automatically migrate VMs from one node to others to assure that the workloads on every node are equal or closer.

The threshold is the average of CPU's usage of all hosts. Based on the threshold, we assign all compute-node into three bands consist of: lightly, moderately and heavily loaded workload.

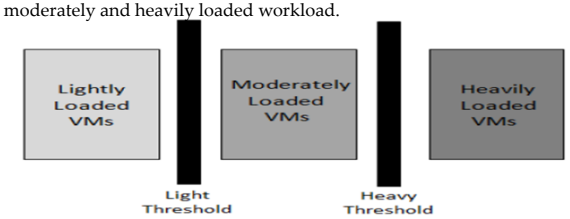

Fig 2: Three regions

Migrate Policy: After calculating the moderately loaded band, the system won't change if all PMs belong to the moderately loaded band. Otherwise, the migration process is triggered when existing at least 2 VMs, one belongs to the lightly loaded band, and the other belongs to the heavily loaded band. A dynamic load balancing is applied to choose which VMs from the overloaded PMs to migrate and which underutilized PMs will receive an incoming migrated data. Figure 3 illustrates the processes of optimizing CPU's usage of three servers.

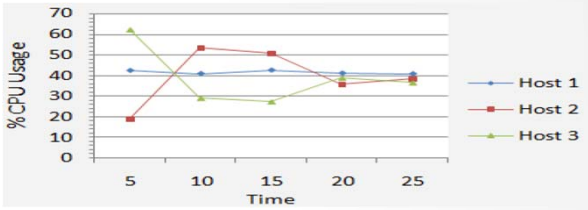

Fig 3: Optimize CPU's usage

C. Resource provisioning
In this part, the load balancing algorithm will handle two extreme scenarios: One, for some reasons some VMs of a PMs suddenly receive a heavy workload lead to rising quickly CPU's usage of this PM up to the high performance and stays there for a period of time. The other scenario is when the head-node distributes the VM requests to a busy cloud cluster when all PMs consumes almost all of its capacity. It means that head-node is unable to find a new PM to distribute the request then the system has to expand their cluster to assume the quality of service.

FUTURE WORK

Load balancing processes guarantee the quality of server but it does not assure the green computing concept. Load balancing leads to consuming a huge amount of energy consumption. Wasting energy could happen when the requests decrease gradually. Next step in green cloud, I'll focus on load consolidation, a process which migrates all virtual machines between a certain amount of physical compute nodes as much as possible and set idle compute nodes in power saving state. Hence, idle machines consume less electricity then generate less heat. The energy consumption of cooling systems reduce then energy consumption of entire system will be reduced. Consolidation process bases on live migration, a technique allowing migrating running virtual machines without rebooting operating system inside it.

REFERENCES

[1] Qian. L., Luo. Z., Du. Y., Guo. L. Cloud computing: A review. 628 – 631.
[2] Raza. K., Patle. V. K., Arya. S. A review on Green Computing for Eco-Friendly and Sustainable IT
[3] Telkikar. S, Talele. S, Vanarse. S. Efficient load balancing using VM migration by QEMU-KVM. *I*nternational *J*ournal of *C*omputer *S*cience and *T*elecommunications. 49 - 53

WKU Ogden Science & Engineering
Department of Computer Science