Computers and Electrical Engineering 47 (2015) 241–254

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Use of proactive and reactive hotspot detection technique to reduce the number of virtual machine migration and energy consumption in cloud data center ☆

Subhadra Bose Shaw *, Anil Kumar Singh

CSED, MNNIT, Allahabad, India

ARTICLE INFO

Article history:
Received 24 March 2015
Received in revised form 22 July 2015
Accepted 23 July 2015
Available online 6 August 2015

Keywords:
Cloud computing
Energy efficiency
Hotspot mitigation
Load prediction
VM consolidation
VM migration

ABSTRACT

The increasing demand of cloud computing motivates the researchers to make cloud environment more efficient for its users and more profitable for the providers. Though virtualization technology helps to increase the resource utilization, still the operational cost of cloud gradually increases mainly due to the consumption of large amount of electrical energy. So to reduce the energy consumption virtual machines (VM) are dynamically consolidated to lesser number of physical machines (PMs) by live VM migration technique. But this may cause SLA violation and the provider is penalized. So to maintain an energy-performance trade-off, the number of VM migration should be minimized. VM migration primarily takes place in two cases: for hotspot mitigation and to switch off the underutilized nodes by migrating all its VMs. If a host is found to be overloaded then instead of immediately migrating some of its VMs we can check whether the migration is really required or not. For this we have proposed a load prediction algorithm to decide whether the migration will be performed or not. After the decision has been taken the algorithm finds a suitable destination host where the VM will be shifted. For this we have proposed a novel approach to decide whether a particular host is suitable as destination depending on its probable future load. We have simulated our algorithms in CloudSim using real world workload traces and compared them with the existing benchmark algorithms. Results show that the proposed methods significantly reduce the number of VM migration and subsequent energy consumption while maintaining the SLA.

1. Introduction

As defined by NIST [1] "*Cloud computing is a model for enabling ubiquitous, 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*". Lack of upfront capital investment, elasticity of resource provisioning and the pay-as-you-go pricing model are continuously attracting people toward cloud computing. Use of multi-tenant model increases the resource utilization. Multi-tenancy allows sharing the same resource among multiple customers in an isolated manner. As cloud can be accessed anytime and anywhere through commodity

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. Danielo Gomes.

* Corresponding author.

E-mail addresses: subhadra.shaw@gmail.com (S.B. Shaw), asbhadoria@gmail.com (A.K. Singh).

http://dx.doi.org/10.1016/j.compeleceng.2015.07.020
0045-7906/© 2015 Elsevier Ltd. All rights reserved.

hardware, its demand is increasing day by day. So it must provide high performance gain to the user and at the same time must be beneficial for the Cloud Service Provider (CSP).

Virtualization is the key technology behind cloud computing that allows the simultaneous execution of diverse tasks over a shared hardware platform. It provides the potential for on-the-fly and on-demand configuration of physical machines to run diverse tasks, hence avoiding resource waste [2]. Cloud provides computing resources in the form of virtual machine, which is an abstract machine that runs on physical machine [3]. Typically a software layer called a hypervisor (e.g. Xen, KVM etc.) or virtual machine monitor (VMM) that resides above the hardware maps the virtual machine to physical resources. Each of the VMs executes either unmodified (full virtualization) or little modified (para-virtualization) version of operating system. Hypervisors support different functions for the hosted VMs such as creation, deletion, restart, suspend and migration [4].

The mapping between VMs and PMs can be changed while applications are running by VM live migration technique [5]. It transfers state of a VM from one physical machine to another with minimum downtime. Three main techniques used for live migration are suspend-and-copy, pre-copy and post-copy. Suspend-and-copy, suspends a VM, copies all its pages and then resumes the VM on the target machine [5]. In this method the downtime is proportional to the size of the VM and the available bandwidth. Pre-copy approach first copies the memory state to the destination through a repetitive process and after that its processor state is transferred to the target machine [6]. Whereas post-copy migration transfers a VM's memory contents using demand-paging after its processor state has been sent to the target host. However, there is a delay associated with each migration, comprising of the time required for the VM [2] to stop execution at the current server, move the accompanying data to the new one and initialize the new VM there.

Irrespective of this overhead, VM migration is indispensable as it helps in hotspot mitigation, dynamic VM consolidation and enables uninterrupted maintenance activities [7]. Hotspot is defined as a condition when a host has inadequate resources to meet the performance demands [7]. Detection of hotspot can be done both proactively and reactively. Compare to proactive hotspot detection technique, reactive one is easier to implement. In this technique a host is considered as overloaded when its utilization level goes beyond a certain threshold. Whereas the former method uses forecasting technique to find hosts which may become overloaded in near future. As prevention is always better than cure, the pre-hotspot-detection technique (proactive) will reduce the number of migration more than the post-hotspot-detection technique (reactive). However, due to inefficient distribution of load, more heat is generated by these overloaded servers which in turn increase the cost of cooling system and substantial emission of CO_2 contributing to greenhouse effect [8]. So to reduce the environmental impact and fulfill the Quality of Service (QoS) requirements specified by users via SLA, some of the VMs have to be migrated. On the other hand, in VM consolidation, VMs are migrated to fewer PMs to reduce server sprawl. Studies have found that servers in many existing data centers are often severely underutilized due to over provisioning for the peak demand [9]. It not only leads to poor resource utilization but also increases the operational cost due to higher consumption of energy. It has been stated that even completely idle server consumes about 70% of the peak power [10]. VM consolidation maximizes the number of inactive physical servers by consolidating the VMs on a minimum number of active physical servers. Ideally, due to the static energy consumed by server components (especially the CPUs), the servers must be kept at that utilization level where it is most energy efficient. The highest utilization level must be less than 100% because at this point performance degradation occurs causing more consumption of energy due to longer execution time [11]. The optimum utilization level varies from processor to processor. For simplicity we have considered a dynamic threshold which denotes the upper limit of the CPU utilization depending on the current system load. However, due to the lack of energy-proportionality in modern server hardware, there's a big difference in energy consumption between an idle and suspended server, suspending an inactive server provides another opportunity for saving energy [12].

There is an inherent trade-off between overload avoidance for achieving high performance gain and dynamic consolidation of VMs to save energy. As to reduce overloading of hosts we should keep the utilization of PM low so that it can deal well with future resource needs. But it results in poor resource utilization and the underutilized servers cause more energy consumption. VM migration is the solution to both problems. So the number of VM migrations should be such that it provides an optimum level of performance as well as saves energy. To ensure the future growth of cloud computing an energy-performance trade-off has to be maintained which helps the cloud providers to fulfill the demand of users with less operational costs.

Most of the current research works [7,8,13–19,21] are based on the current load of the system. If a host is found to be overloaded at present then VM migration is initiated immediately [13]. But each VM migration is associated with some performance degradation which in turn increases the SLA violation. As a result each VM migration increases the operational cost. So the problem is to determine when a VM migration should be initiated so that the cost associated with extra energy consumption and SLA violation can be minimized. To achieve this goal we have proposed a load prediction method which decides whether a VM migration should be initiated or not. Our contributions in this paper are as follows:

- Time-series based forecasting methods are used to predict future load of a system. If a server is currently overloaded and the next predicted load is also greater than the dynamic upper threshold then migration will take place.
- The forecasting method is used to predict multiple (n) future load of the system. If the server is currently overloaded and at least k of the predicted load are greater than the dynamic upper threshold then migration will be initiated.