Computers and Electrical Engineering 000 (2017) 1–13

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Interference aware prediction mechanism for auto scaling in cloud☆

K.R. Remesh Babu[a,*], Philip Samuel[b]

[a] Cochin University of Science and Technology & Assistant Professor, Department of Information Technology, Government Engineering College, Idukki, India
[b] Division of Information Technology, School of Engineering, Cochin University of Science & Technology, Kochi, India

ARTICLE INFO

Article history:
Received 30 April 2017
Revised 6 December 2017
Accepted 7 December 2017
Available online xxx

Keywords:
Auto scaling
Cloud computing
Interference
Load balancing
Quality of Service
Resource management

ABSTRACT

Advancements in cloud computing has transformed it into the most promising computing paradigm for business organizations. In a dynamic cloud environment, Virtual Machine (VM) migration is a critical step in resource management, especially in large scale datacenters. Most of the VM migration and load balancing policies are based on power consumption and response time, with little attention on the interferences caused due to VM migrations. Such interferences degrade the overall performance of the system, and consequently violates the service level agreement (SLA) between the customer and the provider. This paper introduces an interference aware prediction mechanism for VM migration, with auto scaling. The automatic scaling policies help to handle sudden load changes with precise prediction and minimum VM migration. The experimental results and comparative analysis show that the proposed model is capable of predicting the interference more accurately. It also provides an optimum threshold range for VM migration.

1. Introduction

Cloud computing became a promising computing IT paradigm due to the advancements in its technology. It offers low cost on-demand services with minimum user intervention. Today, more and more organizations prefer cloud as their computing platform, rather than the traditional storage mechanism. Nowadays, cloud computing utilities like Amazon Web Services (AWS), Google, etc. have become a part of public life and most of them use these services without knowing it. The popularity of cloud is due to its pay-as-you-use model with a minimum cost and assured Quality of Service (QoS).

The backbone of the cloud is virtualization technology that provides abstraction, elasticity, and agility to its large pool of resources such as Virtual Machines (VMs), storage, and connected networks. A cloud presents these diversified resources as a unique one to the customer, and handles resource management in a user friendly manner. The virtualization also makes the cloud smarter by resource consolidation techniques.

The most important aspect of a computing service is user satisfaction and it doesn't depend on whether the service is deployed in a cloud environment or in a non-cloud environment. If the resource management methods are not fine-tuned there will be frequent VM migrations that will adversely influence the system stability. This VM instability will, in turn, affect the overall system performance and quality of services delivered to the users. An optimal resource allocation strategy should consider these factors and mitigate frequent migrations to improve the QoS offered to the customers. Proper

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. L. Bittencourt.
* Corresponding author.
E-mail address: remeshbabu@gecidukki.ac.in (K.R.R. Babu).

https://doi.org/10.1016/j.compeleceng.2017.12.021
0045-7906/© 2017 Elsevier Ltd. All rights reserved.

monitoring and managing Service Level Agreement (SLA) for every service, is the key task in offering required services with guaranteed quality. Resource management techniques like VM migration and load balancing, resource scalability etc. can be used for maintaining QoS. When the computation load increases or the load variation occurs in a physical machine (PM), the VM migration technique is used to redistribute the load among different physical machines, for optimal task scheduling. This will allow the current operation to be performed in another location. Thus, the VM migration technique can be used to maintain the SLA between the customer and the service provider. The migration operations are done in the background, without the user's knowledge and intervention.

Auto scaling is one of the hot features of cloud computing that facilitates resource scalability beyond datacenter boundaries. Scalability increases the performance of cloud eco system in terms of storage, processing power, throughput, and reliability. Resource scalability improves the throughput of the system without rejecting and reducing the input workload. In an ideal condition, there is no performance degradation due to the migration of VMs. But VM migration will cause certain performance delays in the operation, due to interference. An interference prediction mechanism will help to reduce VM interference and the corresponding performance delay. This paper proposes an interference prediction technique for VM migration that will help in the respective auto scaling of resources. The proposed work is intended for the scalability of resources, when the user workload increases beyond a certain threshold value. So, VMs in a particular host can be migrated to appropriate destinations based on least interference values, for the performance improvement of entire cloud system. This will reduce the number of migrations in the cloud system.

In the dynamic cloud environment, autoscaling of resources enable cloud service providers (CSP) to satisfy customers for their computation needs, without affecting performance. When a particular PM or a CSP itself can't cope with the user requirements, the resources are automatically scaled out to another PM or any CSP. This scaling of resources must be done as fast as possible, since any delay during the execution creates degradation in the performance. The autoscaling should be done based on some already defined threshold values [1]. When workload increases beyond the value of this defined threshold, scaling up has to occur so as to reduce SLA violations. The system must also have to release unused resources by either scale-down or scale-in process, when the workload decreases.

If the system can predict the interferences due to overloaded conditions, the suitable scaling decisions can be taken in advance to ensure performance. This ensures seamless execution of the scheduled user tasks. The proposed prediction model will calculate the interferences more precisely so that it will be easy to scale VM and hence maintain guaranteed SLA. The objective of this work is to make a seamless task execution during the VM migration, using a prediction model in the dynamic cloud environment with ensured SLA, and least prediction error.

The remaining part of this paper is organized as follows. The previous works in this area are reviewed in Section 2. Sections 3 and 4 contains detailed description of the proposed system and Pareto dominant prediction method respectively. The resulting analysis is given in Section 5 and the work is concluded in Section 6.

2. Related works

The researchers are working towards the enhancement of existing live migration technology. Most of the research works focus only on power-saving, cost reduction, and load balancing scenarios. The load balancing algorithms avoid overloaded, underloaded, or idle situations of PMs for better performance, while cost aware methods aim at the reduction of power utilization.

In live-migration, load balancing can be done without losing state information of VMs, so that migration will not affect the performance of cloud datacenter. Recent live migration techniques and opportunities for performance improvements are discussed in [2]. Pre-copy, post-copy, and hybrid are the available different live migration techniques usually adopted in distributed computing for seamless execution.

VM bandwidth allocation can be done in a fair way using an algorithm called Falloc [3]. Here, the fairness is ensured by considering two main objectives:

(i) Guaranteed bandwidth for maintaining SLA.
(ii) Sharing of residual bandwidth.

The algorithm allocates a particular portion of the bandwidth on congested links to VMs, based on bandwidth requirements of the applications. To allocate bandwidth, different weights are assigned for distinct applications. The iAware [4] is a novel live VM placement method. It calculates the interference in PMs based on an empirical formula and the resource demand of VM. The experiments are validated through realistic benchmark workloads on Xen cluster. It is based on demand-supply model and tries to minimize VM migration time and interferences in servers. Another advantage is that it can co-exist with the existing VM scheduling or consolidation algorithms. The algorithm also tries to improve power consumption with load balancing. The final VM placement decision is based on a simple ranking method.

Multiple task execution creates interferences in the system [5]. This article presents a 4-dimensional multiple resource model, along with a brief description about commonly happening interferences during multiple tasks. VMFlocks is an incrementally scalable high performance VM migration service designed for cross datacenter [6]. It efficiently uses the available cloud resources to accelerate data de-duplication and to transfer processes with a minimum access control.

The server consolidation and VM scale-in process create a significant variation in the power efficiency and thermal performance of distributed systems [7]. The contention of resources impacts the distributed system throughput differently,

Please cite this article as: K.R.R. Babu, P. Samuel, Interference aware prediction mechanism for auto scaling in cloud, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.12.021