Computers and Electrical Engineering 000 (2017) 1–16

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Horizontal scaling and aggregation across heterogeneous clouds for resource provisioning☆

Kirthica S.*, Rajeswari Sridhar

Department of Computer Science & Engineering, Anna University, Chennai 600 025, India

ARTICLE INFO

Article history:
Received 21 April 2017
Revised 22 December 2017
Accepted 22 December 2017
Available online xxx

Keywords:
Cloud computing
Resource provisioning
Elasticity
Virtual machine
Cloud inter-operability
Inter-cloud
Multi-cloud

ABSTRACT

A cloud's elasticity is enhanced by inter-operating with other clouds and borrowing resources from them. Hence, a request from any such inter-operating cloud's user is fulfilled with resources of some individual cloud in the network. However, the request is denied on unavailability of sufficient resources in any individual cloud. In this work, this denial of a user's request is prevented by obtaining resources from multiple clouds and hence, the probability of satisfying the request is improved. This proposal intends to replace the existing resource provisioning technique in an open framework, Cloud Inter-operation Toolkit (*CIT*), to achieve an increased transaction success rate which is evident from the experimental results obtained from a real-time heterogeneous cloud environment set up using Eucalyptus, OpenNebula and OpenStack. In spite of satisfying a request with resources from multiple clouds, these resources are aggregated and provided in a way that is easily accessible by the user.

1. Introduction

The one feature of cloud that makes it appealing to its users is *Elasticity* which increases the availability of resources in the cloud [1]. The *resource provisioning technique* used in a cloud is said to be sound if it enhances the cloud's elasticity to the maximum limit. This limit is determined by the amount of physical resources in the cloud.

To overcome the limitation in elasticity of a cloud (say, home cloud), *cloud inter-operation* is resorted to, thereby allowing it to borrow and lend resources across clouds in the same network [2]. In such an inter-operating cloud environment, all other clouds in the network are considered external to the home cloud and hence referred to as *External Clouds* (*ECs*) to that cloud.

Higher the number of requests satisfied by a cloud, higher will be its user satisfaction and higher the probability of users approaching the cloud, thereby resulting in an increased profit to the cloud. Users will be more attracted to the cloud if the response time of the cloud (turnaround time experienced by the user) is low. To achieve this, the cloud should adopt a resource provisioning technique that produces a high transaction success rate that should not be compromised for a lower turnaround time. Hence, trade-offs between transaction success rate and turnaround time are solved by giving priority to the former. Thus, this work aims at increasing the transaction success rate by keeping the turnaround time as low as possible.

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

* Corresponding author.
E-mail addresses: s.kirthica@gmail.com (Kirthica S.), rajisridhar@gmail.com (R. Sridhar).

https://doi.org/10.1016/j.compeleceng.2017.12.044
0045-7906/

A Cloud Inter-operation Toolkit (*CIT*) [3] is an open framework taking care of all activities of the cloud in which it is installed i.e. the home cloud, which is referred to as *CIT cloud* in the rest of the paper. In *CIT*, a technique was proposed to provide resources without compromising on the *CIT cloud's* security while inter-operating with other clouds in the network. This technique produced an increased transaction success rate by borrowing resources requested by a user on insufficiency in the *CIT cloud* from one of the best available clouds in the network.

As a high transaction success rate is guaranteed by borrowing resources from just one other cloud to satisfy a single request, the impact of borrowing from multiple clouds for the same must be better. Kindled by this line of thought, a technique is designed to utilize the combined resources obtained by horizontal scaling across the several clouds in a network to satisfy any request. The crux is that if no single cloud in the same network as the *CIT cloud* has all the resources requested by a user of the *CIT cloud*, there are possibilities that all the clouds in the network will have some amount of resources which when put together can satisfy this single request. Since *CIT* eases inter-operation in a multi-cloud [4] environment, where all other clouds in the network do not volunteer in inter-operating with each other, consolidating resources from these clouds to satisfy a request is a challenging task. In this work, a novel technique of acquiring resources by horizontal scaling across clouds in a multi-cloud environment is designed to achieve an improved transaction success rate.

The resources provided by *CIT* to the users are in the form of virtual machines (VMs), thus targeting IaaS (Infrastructure-as-a-Service). To form a stand-alone VM, minimum quantities of certain essential utilities (like CPU, memory) are required. Hence, for acquiring resources from a cloud, a request should be placed with the minimum quantities of these utilities.

In this work, a new resource provisioning algorithm, coined *Equi Placer* and designed for *CIT* to achieve an improved transaction success rate, introduces the novel concept of horizontal scaling across clouds in a multi-cloud to satisfy a single request. This algorithm adopts 'divide and conquer' strategy to divide an arriving request equally into several small *sub-requests* in such a way that each sub-request has the minimum quantities of all the essential utilities. Hence, each sub-request is satisfied in the form of a VM by any individual cloud in the network (*CIT cloud* inclusive).

The resulting multiple VMs for a single request, if delivered as such, would not ease their accessibility by the user. Hence, a novel idea of *virtual VM*, that gives an illusion of a single VM to the user, is conceptualised in this paper, for delivering the VMs virtually as a single VM to the user, and for thus enabling seamless access of resources. This virtual VM is formed by a tree structure, which aids in a single point of access to all the VMs it comprises, and also, in reducing the access time to all VMs in it. All independent events in the entire process are parallelized to maintain a low turnaround time.

To summarize, the novel concepts introduced in this work are as follows:

- horizontal scaling across clouds in a multi-cloud to satisfy a single request
- virtual VM to aggregate the multiple VMs created for the single request

The rest of this paper is organized as follows: Section 2 examines the existing works on resource provisioning in an inter-operating cloud environment. An overview of *CIT* and the technique used in it to provide resources is given in Section 3. Sections 4 and 5 discuss the details of the proposed algorithms. These algorithms are evaluated and analysed in Section 6. Section 7 concludes the paper with future directions.

2. Prior works

Resources of a cloud are dynamically provided to its users in the form of VMs without overloading any individual server forming the cloud and also utilizing minimum number of servers as possible [5]. Several other related research [6,7] discusses the allocation of VMs in a load balanced fashion within a cloud, thereby trying to optimally utilize its resources distributed across different servers and datacenters.

RESERVOIR [8] provides infrastructure of multiple clouds in a federation [4] to different applications by splitting each application into small independent components each of which runs on the same or different clouds in the federation. The mOSAIC API [9] also divides an application into several components but these components are dependent on each other for the proper functioning of the application. In mOSAIC, based on the usage of the application, the infrastructure provided from a pre-defined set of clouds is scaled up or down. OPTIMIS [10] and Service Superscalar (ServiceSs) [11] also provide infrastructure of cloud(s) to applications with a major focus on the life cycle of the applications. Zeng et al. [12] have handled resource provisioning from Cloud-of-Clouds to applications involving continuous writing and processing. Few other works [13–15] mediate between users and multiple clouds to ease the users' process of selecting one application out of several provided by these clouds.

In Contrail [16] and the dynamic collaborative cloud [17], resources of a cloud federation are provided to users through a single point of contact leading to a bottle-neck. While one of the clouds in the federation is the point of contact in dynamic collaborative cloud, Contrail is a separate entity acting as a broker between the cloud federation and the users. Yet another work, IntCloudWare [18] introduces a broker for establishing the federation and all clouds in it can be queried directly by their users for resources.

STRATOS [19] is an exclusive broker service for providing resources from multiple clouds to applications based on users' needs. Another cloud brokering mechanism, proposed by Tordsson et al. [20], creates a static set of VMs in multiple clouds based on the OpenNebula virtual infrastructure manager [21]. InterCloud [22] consolidates different utilities separately from all clouds in a federation and provides them to the users via a two-tier brokering mechanism: one at the client side and the

Please cite this article as: Kirthica S., R. Sridhar, Horizontal scaling and aggregation across heterogeneous clouds for resource provisioning, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.12.044