Contents lists available at ScienceDirect
International Journal of Approximate Reasoning
www.elsevier.com/locate/ijar

A residue-based approach for resource provisioning by horizontal scaling across heterogeneous clouds ☆

S. Kirthica *, Rajeswari Sridhar
Department of Computer Science and Engineering, College of Engineering, Guindy, Anna University, Chennai 600025, India

A R T I C L E I N F O

Article history:
Received 12 December 2017
Received in revised form 1 June 2018
Accepted 1 July 2018
Available online xxxx

Keywords:
Cloud computing
Resource provisioning
Elasticity
Inter-cloud
Multi-cloud
Cloud ranking

A B S T R A C T

A cloud's elasticity which makes it attractive is limited to its physical resources. As a measure to prevent this limitation from being a hindrance to the attraction, the cloud inter-operates with other clouds. This improves its resource strength by borrowing necessary resources from the clouds with which it inter-operates. Normally, this borrowing happens from one cloud to satisfy one request arriving to it, thereby enhancing its elasticity to a certain extent. Further enhancement is achieved by horizontal scaling across clouds to satisfy a single request. In this work, the enhancement is taken one step further to propose a novel residue-based resource provisioning technique for a cloud to perform horizontal scaling by splitting a request dynamically based on the resource availability in itself and in the clouds with which it inter-operates. Also, a greedy technique is presented to rank clouds and hence, additionally contribute to the enhancement of elasticity and, in turn, to the effectiveness of the resource provisioning technique used in inter-operating cloud environments. Results obtained from real-time heterogeneous cloud environments prove the effectiveness of the proposed algorithms.

1. Introduction

A cloud environment uses one of its features, *Elasticity*, for providing resources on users' demand [1,2]. The effectiveness of elasticity is determined by the resource provisioning technique that enhances it to the maximum possible extent. The performance of the technique is measured by determining the satisfaction rate of the cloud's users and is hence proportional to the *transaction success rate*. Further, *turnaround time* will also greatly influence the user's satisfaction. Although it is not a governing factor for the user's desirability, it is advisable to keep it low.

A cloud's elasticity is limited to its available physical resources. This is overcome by *inter-operating with other clouds* [3,4], thereby enhancing every involved clouds' elasticity and benefiting each other by an increased transaction success rate. In addition, the resources of all these clouds are effectively utilized. In such an inter-operating cloud environment, all clouds external to a cloud are termed *External Clouds* (*ECs*) [5,6] to that cloud. The Cloud Inter-operation Toolkit (*CIT*) [5], as the name suggests, when installed in a cloud helps it in securely inter-operating with a *dynamic* set of clouds notably without the use of any third-party, making it the best cloud inter-operation facilitator.

☆ This paper is part of the Virtual special issue on 1st International Workshop on Uncertainty in Cloud Computing – DEXA 2017, Edited by Allel Hadjali, Haithem MEZNI and Sabeur ARIDHI.

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

https://doi.org/10.1016/j.ijar.2018.07.002
0888-613X/

In general, on resource insufficiency in a cloud, a user's resource request to it is satisfied with resources of any one of its EC. Hence, resources of an EC, if sufficient to satisfy the request of the cloud, aid in increasing the transaction success rate of the cloud. Such a resource provisioning technique is proposed in *CIT* [5], wherein a proof for this increased transaction success rate is available. Also, the turnaround time is maintained as low as possible here.

To satisfy one request, if the resources of just one cloud guarantees an increased transaction success rate, the resources from multiple clouds would give better results. This signifies that horizontal scaling, if done across clouds to satisfy a single request, would aid in further improving the transaction success rate. The aforesaid horizontal scaling is incorporated in a work proposed earlier [6] which indeed produces a higher transaction success rate compared to that in *CIT* [5]. Nevertheless, the turnaround time in the former work is almost thrice that in the latter. Also, horizontal scaling across clouds is achieved by simply dividing a user's request equally into multiple smaller requests which are further indivisible. Clouds whose resources are not sufficient to provide for one of these smaller requests will not be able to contribute to any of the other requests because all these requests are equal. Hence, resources though available in these clouds are not exploited for the request. This is because the resource request is split statically without considering the availability of resources in the inter-operating clouds.

In this work, a novel resource provisioning technique is proposed to improve the transaction success rate even more while maintaining a low turnaround time. Here, to satisfy a request, the clouds participating in the inter-operation are queried one at a time and the maximum amount of resources that each cloud can provide is obtained to fulfil part of the request. The *residue* left unfulfilled by the queried cloud is carried over and the next cloud in row is queried for it. This querying for resources from clouds continues until there is no residue left to fulfil. Thus, horizontal scaling across clouds is achieved to satisfy a single request by splitting the request on the fly based on the resource availability in the inter-operating clouds. Also, this technique is tailored in such a way that satisfaction of SLA (Service Level Agreement) and provisioning of resources do not hinder each other, resulting in a sound performance in both respects. Further, a greedy approach is presented for ranking the participating clouds based on their resource availability to complement any resource provisioning technique in inter-cloud environments.

The remainder of this paper is organized as follows: Section 2 briefs the existing resource provisioning techniques in inter-operating cloud environments. Section 3 gives a brief preface for the proposal. The proposed novel works on resource provisioning in an inter-operating cloud environment and ranking of ECs are respectively detailed in Sections 4 and 5. Section 7 analyses the performance of the proposed works. Section 8 concludes this paper and gives suggestions on future directions.

2. Related work

Inter-operating cloud environments are broadly classified into two major categories [7]: Federated and Multi-cloud. A federated cloud environment is one in which the participating clouds serve their users in an aggregated manner, thereby handling requests and catering resources through a single point of contact. Also, a centralized control exists to monitor the activities of all the clouds. On the contrary, a multi-cloud environment allows the participating clouds to make their own administrative decisions. Apart from participating in the inter-operation, the clouds can also have their own techniques for resource provisioning to satisfy their users.

RESERVOIR [3,8,9] is a federated cloud environment, not involving mediators for resource provisioning, where resources of all the participating clouds are divided to accommodate multiple software application components that are independent of each other for processing. Each component of the application executes using the resources of any division of any cloud in the federation. Dynamic Collaborative Cloud [10,11] is also a federated environment that does not use any broker. Here, one of the clouds is chosen as a primary cloud and the remaining clouds collaborate with it and increases the resource availability in the primary cloud.

InterCloud by Buyya et al. [12,13] utilizes a central third party component for resource organization of all clouds in the federation and also for distribution of these resources to a second level of brokers who in turn deliver the resources to users. Intercloud by Bernstein et al. [14–16], on the other hand, maintains a global repository for information on resource availability in all clouds. Replicas of this repository, made available to brokers chosen by different communities, assist in negotiation between clouds and resource distribution to users.

Federated Cloud Management (FCM) [17–19], as the name suggests, also constructs a federated cloud environment by installing all the clouds with a software component that assists them with resource distribution and other administrative activities. In addition, a broker mediates between the software components in the clouds and the users, and hence assists in resource allocation.

Contrail [20–22] is a central software entity that builds a federated cloud environment and acts a bridge between the federation thus formed and the users. It monitors the activities of all clouds in the federation providing each of them a unique identity. The Unified Framework for data management [23,24] is, as the name suggests, a framework built for data archival in a federated cloud environment. OPTIMIS [25–27] is a toolkit facilitating both federated and multi-cloud environments. This toolkit consists of two parts, one of which is required by the software applications for acquiring resources and the other by the clouds for enabling inter-operation and resource provisioning.

mOSAIC [28–30] is an open-source API for constructing a multi-cloud environment and computing thus enabled is termed Sky Computing [31–33]. Here, a broker is employed for resource distribution to applications which are divided into

Please cite this article in press as: S. Kirthica, R. Sridhar, A residue-based approach for resource provisioning by horizontal scaling across heterogeneous clouds, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.07.002