Accepted Manuscript

A hierarchical structure for optimal resource allocation in geographically distributed clouds

Hasan Ziafat, Seyed Morteza Babamir

PII: S0167-739X(17)32884-4
DOI: https://doi.org/10.1016/j.future.2018.08.027
Reference: FUTURE 4411

To appear in: *Future Generation Computer Systems*

Received date : 16 December 2017
Revised date : 16 August 2018
Accepted date : 16 August 2018

Please cite this article as: H. Ziafat, S.M. Babamir, A hierarchical structure for optimal resource allocation in geographically distributed clouds, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.08.027

A Hierarchical Structure for Optimal Resource Allocation in Geographically Distributed Clouds

Hasan Ziafat[a], Seyed Morteza Babamir[a,*]

[a] *Department of Software Engineering, University of Kashan, Kashan, Iran.*

Abstract

Due to the growth of users' requests for various resources in cloud computing, the optimal resource allocation is one of the most important challenges in cloud environments. The optimal resource allocation is achieved by considering user requirements stated in Service Level Agreements (SLAs) and the Quality of Services (QoSs) provided by resources. Since some user requirements (objectives) conflict with some others, an optimal trade-off between them is required in the selection of resources. Obtaining such a trade-off is a complicated and NP-hard problem because we may come up with a lot of permutations (choices) of the resources with the desired QoS. If a cloud environment is geographically distributed, the problem becomes more complicated because in the geographically distributed cloud there are a lot of candidate datacenters with qualified resources. The user requirements considered in this paper are *availability* and *reliability* of resources should be *maximized* and resource *cost* and *response time* should be *minimized* as well as the *minimization* of the *network traffic*. The maximization of and the minimization of the requirements conflict with each other; therefore a trade-off between them is required. In this paper, a hierarchical structure with two resource selections methods called Simplex Linear Programming (SLP) Method and GrEA-based method are used where the hierarchical structure is used to present connections between distributed datacenters and the methods are used to select optimal resources among the datacenters. The most important feature of the hierarchical structure is to prevent the occurrence of an accumulation of requests in a datacenter leading to increase the rate of finding optimal VMs. Moreover, in most studies on geographically distributed clouds, one user requirement is considered for the optimal selection of resources; however, in this study, datacenters are selected based on 4 user requirements as well as the network traffic. Moreover, while in most studies, the requests for one user is considered at one time, this work does the requests for several users sent to the cloud simultaneously. Our results showed an optimal trade-off between user conflicting requirements and the resource utilization by 92%. Moreover, the performance indexes *coverage ratio* and the *maximum spread* showed more performance compared to related studies.

Keywords: Geographically distributed clouds, Many-objective algorithm, VM allocation, Linear Programming (LP), GrEA algorithm, Hierarchical structure.

1. Introduction

Nowadays, large cloud service providers offer various cloud services with specific Quality of Services (QoSs) to users by establishing numerous geographically distributed datacenters [1], which leads to the improvement of performance [2] and higher reliability, lower costs [3] and lower communication latencies than traditional datacenters. However, the most important challenges for cloud providers are the management of resources by considering (1) the satisfaction of user requirements (objectives) and (2) the restriction of their resources, which leads to an important impact on the utilization of the providers and the user satisfaction. This challenge is more complex in geographically distributed clouds than centralized ones due to the small size and the large number of datacenters leading to the significant communication latency. Moreover, in geographically distributed clouds, more than one datacenter is sometimes required to respond to a user request because one datacenter may have not the sufficient number of resources for a request. Therefore, the allocator must consider a number of datacenters distributed geographically.

*Corresponding Author
Email addresses: Ms.ziafat@gmail.com (Hasan Ziafat), babamir@kashanu.ac.ir (Seyed Morteza Babamir)