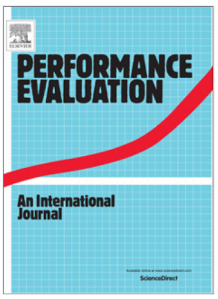

Dynamic placement of resources in cloud computing and network applications

Yuval Rochman, Hanoch Levy, Eli Brosh

PII: S0166-5316(16)30218-8
DOI: http://dx.doi.org/10.1016/j.peva.2017.06.003
Reference: PEVA 1913

To appear in: *Performance Evaluation*

Please cite this article as: Y. Rochman, H. Levy, E. Brosh, Dynamic placement of resources in cloud computing and network applications, *Performance Evaluation* (2017), http://dx.doi.org/10.1016/j.peva.2017.06.003

Dynamic placement of resources in Cloud Computing and Network Applications

Yuval Rochman[a,*], Hanoch Levy[a], Eli Brosh[b]

[a]*School of Computer Science, Tel Aviv University, Ramat Aviv, Tel Aviv 69978 Tel Aviv, Israel*

[b]*Nexar, Ltd.*

Abstract

We address the problem of dynamic resource placement in general networking and cloud computing applications. We consider a large-scale system faced by time varying and regionally distributed demands for various resources. The system operator aims at placing the resources across regions to maximize revenues, and thus needs to address the problem of how to dynamically reposition the resources in reaction to the time varying demand.

The challenge posed by this setting is to deal with arbitrary multi-dimensional stochastic demands which vary over time. Under such settings one should provide a tradeoff between optimizing the resource placement as to meet its demand, and minimizing the number of added and removed resources to the placement. Our analysis and simulations reveal that optimizing the resource placement may inflict huge resource repositioning costs, even if the demand has small fluctuations. We therefore propose an algorithmic framework that overcomes this difficulty and yields very efficient dynamic placements with bounded repositioning costs. Our solution is developed under a very wide cost model, and thus allows accommodation of many systems. Our solutions are based on new analytic techniques utilizing graph theory methodologies that can be applied to other optimization/combinatorial problems.

Keywords: Resource-placement, stochastic, distributed-cloud, graph algorithms.

1. Introduction

Cloud computing has emerged as an attractive solution for building large scale services and geographically distributed applications over the Internet. Popular cloud computing platforms like Amazon EC2 [1] and Microsoft Azure [2] organize a shared pool of (virtual) servers [1] in geographically distributed data-centers to enable on-demand delivery of computer resources at scale. By using a distributed cloud platform, the service provider can place server resources at geographical areas close to its users to provide adequate level of service quality, e.g., low response times, and for better resiliency.

To engineer such a system, the service provider needs to balance two main factors: (a) the revenue from serving a demand, where it is typically better to serve a demand by a resource located in the same area rather than by one located remotely; (b) the cost of placing a resource, which reflects the cost of renting a

*Corresponding author

Email addresses: yuvalroc@gmail.com (Yuval Rochman), hanoch@cs.tau.ac.il (Hanoch Levy)

[1] Virtual servers are also called instances or virtual machines in the cloud computing community.