Computer Networks 93 (2015) 435–447

Contents lists available at ScienceDirect

Computer Networks

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

Elastic virtual machine placement in cloud computing network environments☆

Eleni Kavvadia[a], Spyros Sagiadinos[a], Konstantinos Oikonomou[a,*], Giorgos Tsioutsiouliklis[a], Sonia Aïssa[b]

[a] Ionian University, Department of Informatics, Tsirigoti Square 7, 49100 Corfu, Greece
[b] INRS, University of Quebec, Montreal, QC, Canada

ARTICLE INFO

Article history:
Received 15 February 2015
Revised 7 August 2015
Accepted 19 September 2015
Available online 19 October 2015

Keywords:
Virtual machine
Elastic placement
Cloud computing
Network architecture
Facility location

ABSTRACT

The growth of cloud computing and the need to support the ever increasing number of applications introduces new challenges and gives rise to various optimization problems, such as calculating the number and location of virtual machines instantiating cloud services to minimize a well-defined cost function. This paper introduces a novel cloud computing network architecture that allows for the formulation of the optimization as an Uncapacitated Facility Location (UFL) problem, where a facility corresponds to an instantiation of a particular service (e.g. a virtual machine). Since UFL is not only difficult (NP-hard and requires global information), but also its centralized solution is non-scalable, the approach followed here is distributed and elastic, and relays local information to improve scalability. In particular, virtual machine replication and merging are proposed and analyzed ensuring overall cost reduction. In addition, a policy that employs virtual machine replication and merging along with migration is proposed to reduce the overall cost for using a service. The efficiency of this policy and its limitations are analyzed and discussed, with simulation results supporting the analytical findings and demonstrating a significant overall cost reduction when the proposed policy is implemented.

1. Introduction

Cloud computing has seen significant growth over the last decade allowing for proliferation of numerous applications and gradually changing the distributed networking paradigm to being more centralized, its core being the data center of the cloud service providers [1,2]. As the number of users and cloud services increase, various scalability problems have arisen within data centers, e.g. performance, power efficiency [3], or the network architecture, e.g. bandwidth, time delays [4]. For example, one of the key optimization problems in this area is to find the location of a particular virtual machine that instantiates a cloud service, incorporating the communication costs over the network links, user traffic demands, and maintenance costs for supporting this service, while minimizing the overall cost.

The focus here is this optimization problem from a network perspective. A cloud computing network architecture is introduced that integrates recent trends, e.g. cloud, fog, and access networks [1,4,5]. Virtual machines (or facilities) that instantiate the particular cloud service offered by the corresponding cloud service provider, can be hosted by data centers within the cloud network or by fog devices

☆ Part of this work was originally published in the proceedings of the IEEE International Conference on Communications (ICC), Ottawa, Canada, 10–15 June, 2012.

* Corresponding author. Tel: +30 26610 87708; fax: +30 26610 87766.
E-mail addresses: ekavvadia@ionio.gr (E. Kavvadia), p10sagi@ionio.gr (S. Sagiadinos), okon@ionio.gr (K. Oikonomou), c10tsio@ionio.gr (G. Tsioutsiouliklis), sonia.aissa@ieee.org (S. Aïssa).

http://dx.doi.org/10.1016/j.comnet.2015.09.038
1389-1286/

within the fog network [6,7]. Further details are presented in Section 2.

Cloud computing network architecture allows for formulation of an Uncapacitated Facility Location (UFL) problem [8] to determine the optimal number of virtual machines offering a given service and their location in the network. However, UFL is a large optimization problem, NP-hard and requires global knowledge [8], and for inherently dynamic environments, such as cloud computing networks, existing centralized approaches that solve UFL do not scale. Even approximate approaches require global information, which is prohibitive in this environment. The requirement for a scalable approach is addressed here by exploiting local information in an elastic hill climbing manner that distributedly moves, replicates, and merges service facilities.

Virtual machines at the "Internet as a Service" layer [1] are assumed to instantiate a given service, as further explained in Section 2. Depending on the implementation, a service may consist of multiple components implemented as individual virtual machines. The proposed approach considers all virtual machines that instantiate a service as a single facility, and throughout this paper, service facility (or simply facility) refers to the instantiation of a given service, corresponding to one or more virtual machines. Such a facility is capable of migrating from one location to another, replicate itself and merge with other facilities that instantiate the same service. These operations are allowed only if conditions are satisfied that allow for overall cost reduction.

Specifically, this paper elaborates on facility replication when conditions of overall cost reduction are satisfied, and a facility merging approach is introduced that allows facilities to merge under conditions that ensure overall cost reduction. The previous conditions and the properties of replication and merging are also investigated. A scalable UFL (s-UFL) policy is proposed that integrates facility replication and merging along with facility migration [9] for moving service facilities within the cloud computing network. The efficiency of the proposed policy is studied, showing that cost reduction in cloud computing networks is possible under certain conditions also derived and investigated here. The ease of implementation in the network nodes (i.e., data centers and fog devices) is another advantage of the proposed policy. The main requirement for implementation is a monitoring mechanism to estimate the aggregate incoming/outgoing traffic load of the data center or fog device that hosts a service facility. It should be noted that the problem is formulated here as uncapacitated, even though hardware is never of unlimited capacity. However, the results derived here for the uncapacitated case can easily be extended to apply in capacitated scenarios.

Analytical results are supported by simulation. In particular, the proposed s-UFL policy achieves efficiency by the exploitation of local information resulting in significant cost reduction for supporting cloud services. Various simulation scenarios of more or less than the optimal number of facilities randomly scattered within the cloud computing network, used as the initial setup, reveal the proposed policys behavior.

Cloud computing network architecture is introduced in Section 2 and the corresponding formulation of the UFL problem is presented in Section 3. The analysis of facility replication and merging is presented in Section 4. The proposed s-UFL policy is presented and analyzed in Section 5, and simulation results in Section 6. Previous related work is discussed in Section 7 and conclusions are summarized in Section 8. For ease of readability, the various proofs are included in a separate technical report [10] and the list of important symbols in the Appendix.

2. Cloud computing network architecture

The hardware required to support cloud services is implemented in data centers [11], i.e., plants typically containing thousands of servers in appropriate conditions, e.g. air-conditioned, and suitably interconnected [12]. This hardware is the basis for supporting the wide range of cloud computing services. A typical layered cloud computing architecture [1] that allows for the categorization of cloud services consists of four layers. At the lowest level, the hardware (implemented in data centers) is accessed as a service from the next, infrastructure as a service (IaaS), layer. For example, virtual machines consisting of various levels of computational power, memory, and storage are a service offered at the IaaS layer. The platform as a service (PaaS) layer then abstracts the necessity to deploy applications on virtual machines. Finally, the service as a platform (PaaS) layer offers web based services, multimedia etc., to the end user. More information about the particular layers and their properties can be found in various survey papers, e.g. [1].

Cloud service providers often utilize more than one data center, depending on a variety of criteria, such as location, proximity to users, and energy consumption [13]. A cloud network is formed from a number of data centers as shown in Fig. 1. In this network, the cloud service provider assigns hardware allocated to a particular service. Depending on service type, user demands, topology, etc., the provider may decide to change the service location (i.e., the particular data center hosting it) and/or create more instances of the service and host them accordingly. As stated above, these instances of a service will be referred to as facilities of the given service. For terminology convenience, a facility corresponds to one (or more) virtual machine at the IaaS layer [1].

In some cases, e.g. when there are strict time constraints, a service facility may have to be located close to the end user to be sufficiently supported. Since data centers are expensive and generally cannot be close to the end users, fog computing [4] has been proposed to provide support services close to the end user. Fog computing infrastructure is based on hardware similar to data center infrastructure, but not at the same scale, and so the corresponding hardware is referred to here as a fog device, e.g. [6,7]. A fog device is considered to be capable of hosting a service facility similarly to a data center under the assumption that the hardware requirements will be limited compared to those for a data center. Thus, a fog device follows the layered cloud computing architecture [1]. Depending on the case, it may be a common device, such as a small server, or gateway, etc. [14].

A fog device, along with other such devices, is part of a fog network, as shown in Fig. 1. A fog network falls between the cloud network and the user access network, allowing the user equipment to use the services offered by the cloud service provider, through the network interfaces. Note