Journal of Network and Computer Applications 45 (2014) 35–43

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

On modeling and optimization for composite network–Cloud service provisioning

Jun Huang [a,*], Guoquan Liu [b], Qiang Duan [c]

[a] Department of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[b] Department of Computer Science, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[c] Department of Information Science & Technology, The Pennsylvania State University, Abington, PA 19001, United States

ARTICLE INFO

Article history:
Received 28 August 2013
Received in revised form
22 January 2014
Accepted 15 July 2014
Available online 24 July 2014

Keywords:
Cloud computing
Network-as-a-Service (NaaS)
Modeling
Service selection

ABSTRACT

Networking plays a crucial role in Cloud computing, which calls for a holistic vision of combined control and optimization for both networking and computing resources. Essentially the services received by Cloud users are composite services that consist of not only Cloud computing functions but also networking services. Therefore, composite network–Cloud service provisioning becomes one of the main networking challenges in Cloud computing. In this paper we tackle the problem of modeling and optimization for composite network–Cloud service provisioning. We first present a system model in which a Network-as-a-Service (NaaS) paradigm is given to facilitate convergence of network and Cloud services; then we formulate QoS-aware network–Cloud service selection as a variant of Multi-Constrained Optimal Path (MCOP) problem and propose an exact algorithm for solving this problem. Theoretical analysis shows that the proposed algorithm is light-weighted and cost-effective in selecting services for composite network–Cloud service provisioning. Experimental results reported in this paper also validate the efficiency of the proposed algorithm.

1. Introduction

One of the most significant recent progresses in the field of information technology is Cloud computing, which changes the way people do computing and manage information. Cloud computing can be defined as a large scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable computing functions and services are delivered on demand to external customers (Armbrust et al., 2010). As an X-as-a-Service (XaaS) based model, the Cloud provides computing functions and resources as services. Users exploit Cloud services to save the expenses of purchasing and managing their own computing facilities.

Networking plays a crucial role in Cloud computing. Cloud services are delivered to customers over networks. From a user perspective, Cloud service provisioning consists of not only computing functions provided by the Cloud but also data communications offered by network services. Results obtained from recent performance study of some commercial Clouds, such as Amazon EC2, have indicated that networking performance has a significant impact on the quality of Cloud services, and in many cases networks become the bottleneck that limits Clouds from supporting high-performance applications (Jackson et al., 2010; Wang and Ng, 2010). Networks with Quality of Service (QoS) capabilities become an indispensable ingredient for Cloud service provisioning; therefore networking and Cloud computing systems should be seamlessly integrated into a composite service provisioning system in order to offer high-performance Cloud services.

Networking systems are facing the challenge of rapidly developing and deploying new functions and services for supporting diverse requirements of various computing applications. A promising approach to addressing these challenges is network virtualization, which decouples service provisioning from network infrastructure and exposing underlying network functionalities through resource abstraction. Network virtualization is expected to become a key attribute of the future networking paradigm (Chowdhury and Boutaba, 2010).

Virtualization, as a potential enabler of profound changes in both computing and networking domains, is expected to bridge the gap between these two fields that traditionally live quite apart; thus enabling a convergence of Cloud computing and networking. Network and Cloud convergence allows combined management,

* Corresponding author.

http://dx.doi.org/10.1016/j.jnca.2014.07.009
1084-8045/

control, and optimization of networking as well as computing resources in the Cloud. With the advent of network virtualization Cloud service provisioning may be significantly improved via implementing virtual networks that offer customized networking solutions for Cloud services.

A key requirement for realizing the notion of network–Cloud convergence is an interaction mechanism that enables flexible and effective collaboration between networking and Cloud computing systems. The Service-Oriented Architecture (SOA) facilitates virtualization by encapsulating system resources and provides a loose-coupling interaction mechanism among system modules (Erl, 2005). SOA has been widely adopted in Cloud computing via the paradigms of Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS).

Recent application of SOA in telecom and networking areas has also attracted extensive research interest (Magedanz et al., 2007). Applying SOA in networking allows encapsulation and virtualization of networking resources in the form of SOA-compliant *network services*. This enables a *Network-as-a-Service* (NaaS) paradigm that exposes networking resources and functionalities as services that can be composed with computing services in a Cloud environment. Therefore SOA may facilitate virtualization in both Cloud computing and networking; thus providing a promising approach to enable convergence of networking and Cloud computing (Duan et al., 2012).

From a service provisioning perspective, the services delivered to Cloud users are essentially composite network–Cloud services that comprise of both computing services provided by Cloud infrastructure and communication services offered by network infrastructure. The service performance that a Cloud user experiences is determined by the QoS of both Cloud services and network services. For example, suppose a biology lab creates 50 GB of raw data that will be processed in Amazon EC2 Cloud. Assume that the lab obtained 10 EC2 virtual machine instances and each instance can process 20 GB data per hour, then the total process time of the Cloud service is only 15 min. However, if the lab uses a network service that offers 200 Mb/s throughput for data transmission to the EC2 server, then even the single-trip transmission delay will be 2000 s or more than 33 min. This simple example demonstrates that for data-intensive computing applications, network service QoS plays a dominating role in end-to-end service performance for Cloud computing.

Therefore, one of the networking challenges in Cloud computing lies in modeling and optimization for composite network–Cloud service provisioning. However, current existing models mainly focus on either networking or Cloud computing systems; thus lacking the ability to characterize both. As a result, a comprehensive model to portrait convergent network and Cloud service provisioning is needed. On the other hand, selecting optimal service sequences from network–Cloud service pool while satisfying various QoS requirements is a challenging optimization problem, which is usually formulated as the Multi-Constrained Optimal Path (MCOP) problem that is known to be NP-hard. The research work presented in this paper addresses these two issues. Following the SOA principle, we first present a layered framework for Cloud and network service composition, in which NaaS enables networking resources to be involved in Cloud computing as full participants just like computing resources such as CPU capacities and memory/disk space. Then based on this framework, we propose a system model for composite network–Cloud service provisioning and develop an efficient QoS-aware service selection algorithm for network–Cloud service composition. Note that the proposed algorithm not only optimizes the resource configuration and management, but also optimizes the user's experience on Cloud service provisioning. The research work presented in this paper can be realized as an extension of our previous work reported in Huang et al. (2012); Duan (2011). Specifically we have made the following contributions in this paper.

- We propose a system model for composite network–Cloud service provisioning and formulate QoS-aware network–Cloud service selection as a variant of MCOP.
- We develop an exact algorithm to solve the problem of QoS-aware service selection based on the proposed model. Also, we analyze the theoretical properties of the proposed algorithm and show that the algorithm is light-weighted and cost-effective in selecting services for composite network–Cloud service provisioning.
- We conduct thorough comparisons between our proposed algorithm and a modified version of the best-known MCOP algorithm through numerical experiments. Our results show that the model and algorithms proposed in this paper are general and efficient; thus offering desirable experience for Cloud end users.

The rest of this paper is organized as follows. We first give a brief review of related work in Section 2. Then in Section 3 we present a NaaS-based framework for network–Cloud service composition, propose a model for composite network–Cloud service provisioning, and formulate the problem as a variant of MCOP. In Section 4 we develop an exact algorithm to solve the service selection problem and then present theoretical analysis on the proposed algorithm. Section 5 compares the performance of our algorithm against that of a variant of best-known MCOP algorithm through numerical experiments. We draw conclusions in Section 6.

2. Related work

Service composition has been extensively studied in the field of Web services and Cloud computing. Numerous technologies have been developed to achieve functional and/or performance requirements of Web service composition, most of which are based on either workflow management or AI-planning approach. Surveys of recent research on Web service composition can be found in Rao and Su (2004) and Dustdar and Schreiner (2005). Optimization of multiple QoS criteria in service composition has been modeled as a multi-dimension multi-choice 0–1 knapsack problem and solved by integer programming (Zeng et al., 2004) or by heuristic search methods (Yu et al., 2007). Genetic algorithms offer another approach to address this problem with the advantage of being able to handle nonlinear QoS constraints (Canfora et al., 2005). In Bao and Dou (2012) Cloud services are modeled as finite state machines and a simple additive weighting technique is used to select an optimal service path. More research on QoS-aware Web service composition can be found in the survey presented in Strunk (2010).

The aforementioned research results were developed for Web/Cloud service composition with little consideration of the network services that interconnect atomic service components; therefore, QoS optimization is limited within computing services and the obtained composite services may not achieve optimal end-to-end QoS (including QoS of both networking and computing services). The modeling and optimization approaches proposed in this paper take into account QoS metrics of both network and Cloud services; thus finding a service path that achieves optimal end-to-end QoS performance.

Network-aware service composition in Cloud was studied in Klein et al. (2012). The authors first build a network model for estimating the network latency between arbitrary services and potential users, then develop a selection algorithm that leverages this model to find service compositions that will result in a low latency. In Klein et al. (2012) the QoS of Cloud services and QoS of