Chapter 17
Service Level Provisioning for Cloud-Based Applications

Valeria Cardellini
University of Roma, Italy

Emiliano Casalicchio
University of Roma, Italy

Luca Silvestri
University of Roma, Italy

ABSTRACT

Cloud computing has recently emerged in the landscape of Information Technology as a compelling paradigm for managing and delivering services over the Internet in a performance- and cost-effective way. However, its development is still at its infancy, with many issues worthy to be investigated. In this chapter, we analyze the problem of service level provisioning and the possible strategies that can be used to tackle it at the various layers of the cloud architecture, focusing on the perspective of cloud-based application providers. We also propose an approach for the dynamic QoS provisioning of cloud-based applications which takes into account that the provider has to fulfill the service level settled with the application users while minimizing the resources outsourced from the cloud infrastructure in such a way to maximize its profits.

DOI: 10.4018/978-1-61350-110-8.ch017

INTRODUCTION

Cloud computing has emerged as a new paradigm in IT that lets cloud-service clients deploy their applications in a large-scale environment with an expectation of good scalability, availability, fault tolerance, and reduced administration costs. Cloud computing enables open market service providers and enterprises to outsource computational and storage utilities with the promise to scale up in case of peak load, to rely on high availability, and to drastically reduce the start up and management cost of data centers. Moreover, the "elastic" property of cloud computing, if properly exploited, avoids over provisioning of resources in case of scarce demand that, along with the resource sharing at infrastructure level, contributes to energy saving.

To exploit the opportunities offered by cloud computing, enterprise and service providers need new mechanisms to dynamically and adaptively plan the capacity that can be outsourced from the infrastructures and platforms, thus to provide their end users with the agreed level of service while minimizing the leasing costs they pay to the infrastructures and platforms providers.

The cloud computing architecture consists of three abstract layers: the Infrastructure as a Service (IaaS) layer at the bottom of the stack (*e.g.*, Amazon EC2, Eucaliptus, InterGrid), the Platform as a Service (PaaS) layer collocated in the middle (*e.g.*, Microsoft Azure, Google AppEngine), and the Software/Application as a Service (SaaS) layer at the top that features a complete application offered as a service (*e.g.*, Google Apps, Facebook, YouTube).

The cloud infrastructure owners (that is, the IaaS providers) have to face the problem of virtualized and physical resource management in order to provide the agreed levels of service to their IaaS customers. These problems have been widely addressed in literature, *e.g.*, (Nguyen Van, Dang Tran, & Menaud, 2009), (Song et al., 2009), (Chen, Wo, & Li, 2009) to mention few. In the same way, there is a growing interest in the provisioning of some Quality of Service (QoS) guarantees at the upper layers (*i.e.*, PaaS and SaaS) of the cloud architecture (Nguyen Van, Dang Tran, & Menaud, 2009), (Fakhouri et al., 2001), (Wang et al., 2007) (Urgaonkar et al., 2008), (Lim et al. 2009), (De Assuncao, Di Costanzo, & Buyya, 2009). QoS provisioning is not a new issue in networked and distributed systems; however, cloud and service computing paradigms increase the system complexity and scale, therefore posing new challenges that are worth to be faced.

In this chapter we discuss the problem of service level provisioning from the perspective of the IaaS customers that provide cloud-based applications to end users (Lim et al., 2009). We first classify the QoS provisioning strategies on the basis of the cloud architecture layer at which they can be applied to and on the basis of the entity that can operate these strategies. Then, we discuss the main issues in dynamic QoS provisioning for IaaS customers and the possible solutions. We next propose some QoS provisioning algorithms for cloud-based applications that we have designed and evaluated through simulation experiments. Cloud-based services are characterized by two key features (i.e., pay-per-use and on-demand resource provisioning) whose economic impact is that consumers only pay for the resources they actually use. Therefore, in the cloud computing environment, the cloud service providers have to minimize the usage of the cloud infrastructure while maximizing the resource utilization in such a way to maximize their profits by fulfilling the obligations settled in a Service Level Agreement (SLA) with their consumers. The approach to dynamic service provisioning we present in this chapter aims to achieve these criteria. Specifically, we propose and evaluate through simulation experiments two algorithms. Both algorithms aim to dimension the pool of resources that the provider offering the cloud-based application has to lease from the cloud infrastructure. The first algorithm reacts to occurred SLA violations, while the second