Chapter 5

Application-Level Monitoring and SLA Violation Detection for Multi-Tenant Cloud Services

Vincent C. Emeakaroha
University College Cork, Ireland

Marco A. S. Netto
IBM Research, Brazil

Ivona Brandic
Vienna University of Technology, Austria

César A. F. De Rose
PUCRS, Brazil

ABSTRACT

Keeping the quality of service defined by Service Level Agreements (SLAs) is a key factor to facilitate business operations of Cloud providers. SLA enforcement relies on resource and application monitoring—a topic that has been investigated by various Cloud-related projects. Application-level monitoring still represents an open research issue, especially for billing and accounting purposes. Such a monitoring is becoming fundamental, as Cloud services are multi-tenant, thus having users sharing the same resources. This chapter describes key challenges on application provisioning and SLA enforcement in Clouds, introduces a Cloud Application and SLA monitoring architecture, and proposes two methods for determining the frequency that applications needs to be monitored. The authors evaluate their architecture on a real Cloud testbed using applications that exhibit heterogeneous behaviors. The achieved results show that the architecture is low intrusive, able to monitor resources and applications, detect SLA violations, and automatically suggest effective measurement intervals for various workloads.

1. INTRODUCTION

Cloud computing facilitates on-demand and scalable resource provisioning as services in a pay-as-you-go manner (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009) thereby making resources available at all times from every location. Like in other business engagements, resource and service provisioning in Clouds are based on Service Level Agreements (SLAs), which are contracts signed between providers and their customers detailing the terms of the provisioning including non-functional requirements, such as Quality of Service (QoS) and penalties in case of violations (Comuzzi, Kotsokalis, Spanoudakis, & Yahyapour, 2009; Buyya et al., 2009).
To establish Cloud computing as a reliable state of the art form of on-demand computing, Cloud vendors have to offer scalability, reliable resources, competitive prices, and minimize interactions with the customers in case of failures or environmental changes. However, ensuring SLAs for different Cloud actors at different layers (i.e., resource, platform, and application) is a challenging task, especially for the application layer. Monitoring at this layer is necessary as several applications may share the same VMs (e.g., to reduce energy consumption and cost) or one application may run on multiple Virtual Machines (VMs) (e.g., large scale distributed or parallel applications).

Although a large body of work considers the development of reliable Cloud management infrastructure (Koller & Schubert, 2007; Frutos & Kotsiopoulous, 2009; Boniface, Phillips, Sanchez-Mcian, & Surridge, 2007), there is still a lack of efficient application monitoring infrastructures capable of monitoring and detecting SLA violations of different customer applications that share the same Cloud resources—mostly virtual machines. Since monitoring might be intrusive and drive down the overall performance of the system, determination of the optimal monitoring intervals is another important issue. The application level monitoring is particularly relevant when considering provisioning of multi-tenant Cloud resources and services.

This chapter describes the current challenges on application monitoring in Clouds, introduces the design and implementation details of a Cloud Application and SLA monitoring architecture, and proposes two methods for determining the frequency that applications need to be monitored. Furthermore, it analyzes the cost of missing SLA violation detections from the customer perspective as initially proposed in our previous work (Emeakaroha, Ferreto, Netto, Brandic, & Rose, 2012). The core component of the architecture is the application-level monitor, which is capable of monitoring application metrics at runtime to determine their resource consumption behaviors and performance. The main contributions of this chapter are:

1. Identification and discussion of challenges facing application provisioning and SLA enforcement especially in multi-tenant Cloud services,
2. Design choices and implementation details of the proposed architecture,
3. Empirical evaluation of two methods for optimal monitoring interval determination: one based on a utility model that considers provider’s profit and SLA objectives and another based on the balance of provider’s penalty costs for violating SLAs and customer’s business impact for missing SLA violation detections; and
4. Evaluation of the impact of multi-tenancy on both the architecture monitoring intrusive-ness and provider’s net utility.

The rest of the chapter is organized as follows: Section 2 presents views on application provisioning and SLA enforcement in Clouds. It also discusses the current challenges and research efforts. In Section 3, we describe the proposed architecture and give details of its design and implementation. Section 4 presents evaluations of the architecture functions using three ray-tracing applications representing customers’ services with different SLAs. We also present intrusion analysis using different monitoring frequencies. Section 5 presents further evaluations to determine the effective measurement interval to detect SLA violations in order to increase provider’s profit, and also to make a comparison between the provider’s and customer’s costs. Finally, in Sections 6 and 7, we present the future research directions and conclusions, respectively.

2. BACKGROUND

Cloud computing facilitates efficient service provisioning mechanisms at different layers. The three well-known delivery models for service provisioning in Clouds are Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) as shown in Figure 1.