Future Generation Computer Systems 31 (2014) 1–11

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

SLA-driven dynamic cloud resource management

Andrés García García*, Ignacio Blanquer Espert, Vicente Hernández García

Institut d'Instrumentació per a Imatge Molecular (I3M). Universitat Politècnica de València, Camino de Vera S/N, 46022 Valencia, Spain

HIGHLIGHTS

- A methodology for the representation of Cloud resources.
- A SLA-driven architecture for the automatic scheduling of Cloud resources.
- A SLA-driven architecture for the dynamic management of Cloud resources.
- The resolution of a use case by a prototype implementation.

ARTICLE INFO

Article history:
Received 6 June 2012
Received in revised form
2 October 2013
Accepted 4 October 2013
Available online 15 October 2013

Keywords:
Service Level Agreement
Cloud computing
Quality of service
Monitoring

ABSTRACT

As the size and complexity of Cloud systems increase, the manual management of these solutions becomes a challenging issue as more personnel, resources and expertise are needed. Service Level Agreement (SLA)-aware autonomic cloud solutions enable managing large scale infrastructure management meanwhile supporting multiple dynamic requirement from users. This paper contributes to these topics by the introduction of Cloudcompaas, a SLA-aware PaaS Cloud platform that manages the complete resource lifecycle. This platform features an extension of the SLA specification WS-Agreement, tailored to the specific needs of Cloud Computing. In particular, Cloudcompaas enables Cloud providers with a generic SLA model to deal with higher-level metrics, closer to end-user perception, and with flexible composition of the requirements of multiple actors in the computational scene. Moreover, Cloudcompaas provides a framework for general Cloud computing applications that could be dynamically adapted to correct the QoS violations by using the elasticity features of Cloud infrastructures. The effectiveness of this solution is demonstrated in this paper through a simulation that considers several realistic workload profiles, where Cloudcompaas achieves minimum cost and maximum efficiency, under highly heterogeneous utilization patterns.

1. Introduction

Cloud computing is currently being used in different application domains, such as industry, science, and government [1–3]. Therefore concepts related to Cloud, such as Utility Computing or Service-Oriented Infrastructures (SOI) have increased their popularity and usage. As infrastructure providers have consolidated a mature market it is important to advance in the efficient provisioning of services. However, Cloud standards such as Open Cloud Computing Interface (OCCI), Cloud Infrastructure Management Interface (CIMI), Cloud Data Management Interface (CDMI), etc., have emerged, opening new possibilities for interoperability and federation.

The assurance of Quality of Service (QoS) to the applications, although identified as a key feature since long ago [4], is one of the fundamental problems that remain unsolved. In the context of Cloud Computing, QoS is defined as the measure of the compliance of certain user requirement in the delivery of a Cloud resource. Although users may define their requirements considering low-level metrics such as CPU or memory load for a virtual machine, users are interested in expressing their requirements as more abstract and higher level concepts such as response time or availability for a service. The growth of complexity, size and scope of Cloud solutions makes it difficult to anticipate how the systems will behave, which is a premise for maintaining an acceptable level of QoS. Therefore, several research groups, both from academia and industry, have started working on describing the QoS levels that define the conditions for the service to be delivered, as well as on developing the necessary means to effectively manage and evaluate the status of these conditions.

There is currently a considerable number of applications that can benefit from the support of QoS in the Cloud. Multimedia and real time applications [5] need instantaneous computing power, Healthcare applications [6] need to ensure the integrity of the security chain and the fulfillment of legal issues, and scientific workflows [7] need to adapt to response time deadlines without

* Corresponding author. Tel.: +34 963236111.
E-mail address: angarg12@upv.es (A. García García).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.10.005

exceeding a maximum budget. For example, in the BonFIRE project [8] the use of QoS, specified in application-level terms, leads to an increase in the efficiency perceived by final users [9]. The applications targeted by this study cover a wide variety of algorithms and data access models, including map-reduce, matrix computation, and graph traversal.

Authors of [10] were one of the first to focus attention on the role of Cloud computing to deliver a sustainable, competitive and secure computing utility. They propose Service Level Agreements (SLAs) as the vehicle for the definition of QoS guarantees, and the provision and management of resources. An SLA is a formal contract between providers and consumers, which defines the resources, the QoS, the obligations and the guarantees in the delivery of a specific good. In the context of Cloud computing, SLAs are considered to be machine readable documents, which are automatically managed by the provider's platform.

The main objective of this paper is to advance on the vision of Cloud computing as a utility, providing components for fulfilling the requirements of applications that require QoS guarantees. To this end, this paper presents Cloudcompaas,[1] a SLA-driven Cloud platform that manages the complete lifecycle of the resources through the utilization of agreements. Cloudcompaas covers all the steps involved on the management of SLAs, from the set-up of the agreement with the final user, feeding the agreement into the Cloud provider and interacting with the manager that allocates the required resources in the infrastructure, to the monitoring of the agreement and performing the necessary actions, in order to maintain the quality levels specified in the SLA.

Cloudcompaas is based on standards, such as the WS-Agreement [11] specification, for defining the agreements, and on open-source initiatives, such as the WSAG4J [12] framework, for implementing a proof-of-concept prototype. In this paper, the WS-Agreement specification has been tailored to meet the needs of Cloud computing, and the WSAG4J framework has been extended and adapted to deal with the complete lifecycle of the agreement, as well as with other requirements that are specific to the domain.

The main contributions of this paper are:

(i) proposing SLAs and the WS-Agreement specification as a mean for the representation of Cloud resources. This methodology is illustrated with a concrete and extensible example model of generic Cloud resources;
(ii) proposing a novel SLA-driven architecture for the automatic provision, scheduling, allocation and dynamic management of Cloud resources. This architecture is based on the WS-Agreement specification. It provides a multi-provider and multilevel framework that allow building and deploying arbitrary Cloud services that span different levels of the Cloud. This architecture allows defining arbitrary QoS rules as well as arbitrary preemptive and corrective self-management actions;
(iii) demonstrating the capabilities of the proposed architecture by the resolution of a use case by a prototype implementation. A set of experiments, using real world load profiles show the improvement on performance in terms of cost and number of failed user requests, of the proposed architecture using elasticity (upscaling and downscaling) rules.

The paper is organized as follows. Section 2 provides an overview about related works on the topic of SLAs and Cloud. Section 3 includes a brief description of the WS-Agreement specification and the WSAG4J framework. The Cloudcompaas platform is introduced in Section 4, including a model for resource representation, the Cloudcompaas architecture and its components. Section 5 introduces the Cloud resources management cycle and operations, as well as the major contributions introduced by Cloudcompaas to this field. Section 6 introduces a use scenario on Cloudcompaas platform, completed with a set of experiments and discussion about the experimental results. Finally, Section 7 summarizes the conclusions of the paper.

2. Related works

Earlier definitions propose SLAs as a mean for the definition of QoS constraints in electronic services [13]. Other works [14] propose SLA as a mean for the autonomic management of services. More recently these two concepts were brought together and SLA is used both for the definition of requirements of resources and the automatic management of the complete lifecycle of such resources.

Several specifications exist targeting SLA definition and management, with different levels of maturity and completeness. WS-Agreement is one of the most important specifications, which provides a protocol for establishing an agreement between two parties that is the basis for the agreement definition language and SLA management cycle presented in this paper.

The Web Service Level Agreement [15] (WSLA) is a framework and a specification developed by IBM for the definition and monitoring of SLAs in a machine readable format within the domain of web services. The WSLA language defines the parties involved in the agreement, the description of the service that the provider delivers to the consumer and the obligations of the agreement, where the guarantees and constrains of the SLA are defined. The WSLA framework is a tool for the SLA-driven management of the lifecycle of web services, using the WSLA specification. This framework integrates the usage of WSLA with other web services standards such as WSDL.

Other proposals for the definition of SLA are the SLAng [16] and WSOL [17] languages. Both proposals are XML-based languages whose aim is to define QoS constraints in the domain of web services, and therefore are tightly related to the web services technologies and standards. Unlike the aforementioned proposals, these two specifications only define a language for the expression of QoS levels, but do not account for the agreement lifecycle or assessment.

Similarly, significant advances have been made in the development of SLA-aware distributed computing systems. In particular, several innovative projects have considered SLA-aware automatic resource management in the last decade.

In [18] an architecture for the provisioning of on-demand virtualized services based on SLA is proposed. The authors define it as "the first attempt to combine SLA-based resource negotiations with virtualized resources in terms of on-demand service provision", and represents a first step in the line of automated SLA-aware Clouds systems. Further works deal with specific facets of SLA management, such as a system for the monitoring of low level metrics in distributed environments and its transformation to high level SLA parameters [19].

More recently [20] proposes an architecture for an SLA-oriented resource provisioning model for Cloud Computing. This architecture is realized using the Aneka platform [21], a solution that enables QoS-driven resource provisioning for scientific computations, and provides mechanisms for the definition of deadline constraints and the incorporation of multiple Cloud resources.

Several European projects in the last years are related at different degrees to the SLA-aware management of resources and other topics covered by Cloudcompaas.

Reservoir [22] is a pioneering European project whose aim is to enable providers to build their own virtualized Cloud infrastructures. Although Reservoir does not cover SLAs and dynamic management of resources, a number of spin-out technologies and

[1] http://www.grycap.upv.es/compaas/.