Future Generation Computer Systems 64 (2016) 39–49

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

BECloud: A new approach to analyse elasticity enablers of cloud services

Marta Beltrán

Department of Computing, ETSII, Universidad Rey Juan Carlos, 28933 Mostoles, Madrid, Spain

HIGHLIGHTS

- Elasticity is a key performance figure in cloud computing.
- But there is not a standard metric or procedure to quantify it and it is rarely used.
- We define a new elasticity metric, general, flexible, simple and easy to measure.
- An analysis procedure and a benchmarking tool, BECloud, are also proposed.
- These definitions allow providers and users to analyse service elasticity enablers.

ARTICLE INFO

Article history:
Received 27 December 2015
Received in revised form
3 May 2016
Accepted 11 May 2016
Available online 26 May 2016

Keywords:
Cloud computing
Elasticity
Performance evaluation

ABSTRACT

Elasticity is a key property of cloud computing but there is a lack of standard elasticity metrics or analysis procedures to easily quantify this performance figure of cloud services. This absence of a unique general elasticity metric makes difficult to consider elasticity as a service level objective in Service Level Agreements, to benchmark cloud services or to explicitly improve the elasticity of scaling and provisioning mechanisms, to mention only some examples. This paper defines a new elasticity metric capable of considering its four main components, scalability, accuracy, time and cost, independently of the service level (infrastructure, platform or software). Furthermore, an analysis procedure to evaluate the behaviour of service elasticity and a benchmarking tool to automate this analysis are presented. The main elasticity enablers of cloud services are identified and analysed using this metric, procedure and tool via real use cases on private and public clouds, drawing interesting conclusions about this important performance aspect of cloud services.

1. Introduction

Many cloud computing definitions identify elasticity as one of the main benefits of this paradigm, understanding this elasticity as the capacity of the cloud to automatically scale up or down the capacity leased by one end user responding to his or her requirements and demands. The realization of true utility computing based on services really offering unlimited and/or immediately available resources at different levels must necessarily be a revolution for all kinds of applications. But experience shows us that to fully achieve this benefit, important challenges need to be faced such as resource provisioning, virtualization management, application provisioning, programming and billing models, etc.

In this work a new elasticity metric is defined for cloud environments, allowing both, end users and providers, to easily quantify this aspect with a unique, general, simple, easy to measure and to understand metric. This along with the fact that the proposed metric allows flexibility and personalization through the use of a general *QoS* performance figure function and a deep evaluation of the elasticity of a cloud service using a simple analysis procedure and an automatic benchmarking tool, not only its punctual quantification, increases the chance for the adoption of our approach.

More specifically the main contributions of this work are (a) An analysis of the elasticity concept in cloud environments determining its main components, (b) The definition of a new elasticity metric capable of capturing these components with a keep-it-simple philosophy, (c) The proposition of an elasticity analysis procedure based on this metric and the development of an automatic benchmarking tool, BECloud (*B*uilding *E*lastic *Cloud*), to help end users to perform this analysis, (d) The identification of the main elasticity enablers (the most important factors affecting it) in common cloud environments, and (e) The validation of the

E-mail address: marta.beltran@urjc.es.

http://dx.doi.org/10.1016/j.future.2016.05.014
0167-739X/

proposed approach and the evaluation of these enablers with experiments performed on real environments, a private SaaS cloud and a public IaaS cloud (Amazon EC2).

The rest of this paper is organized as follows. Section 2 analyses and discusses the elasticity concept, the related background and its implications in cloud environments. Section 3 presents the considered context and the problem formulation. The new elasticity metric, the procedure proposed to analyse it and the developed benchmarking tool are defined and described in Sections 4–6 respectively. Section 7 discusses the main elasticity enablers identified in this research, while Section 8 shows and discusses the most important experimental results obtained to validate our approach in real use cases. And finally Section 9 summarizes the conclusions of this work and the most interesting lines for future research.

2. Related work

Some cloud computing definitions make particular emphasis on the importance of elasticity as a basic property of the paradigm, in fact, elasticity definitions emerged together with the cloud paradigm. But these works do not define elasticity metrics to quantify this essential property. In [1] a simple elasticity model is proposed: if $D(t)$ is the resource demand function and $R(t)$ is the provisioned resources function, perfect elasticity is achieved when $D(t) = R(t)\ \forall t$. This model assigns a cost to the situations in which $D(t) > R(t)$ (under-provisioning) and to the opposite situations in which $D(t) < R(t)$ (over-provisioning) and tries to quantify the total cost caused by not having a perfect elasticity.

This approach has been improved later in [2], using more sophisticated models and data obtained with real workloads executing on public clouds to assign costs to under-provisioning and over-provisioning situations. This new model considers a penalty for provisioning resources that are not really needed but also for releasing resources that may be needed again soon. The figure of merit defined to perform comparisons of elasticity among different cloud services is related to these penalties. In [3] an elasticity model is proposed too, in this case in order to understand the elasticity requirements of a given application and if the elasticity provided by a cloud provider is able to meet those requirements. This work is focused on evaluating and comparing IaaS providers. Again considering IaaS providers, [4] is focused on defining an elasticity economics platform to allow cloud users to evaluate economic aspects of different elasticity rules and to perform trade-off analysis of different cost-performance metrics under varying workload patterns.

On one hand, these definitions and models do not provide a general elasticity metric definition, but they all identify the aforementioned components (scalability, accuracy, time and cost), quantifying, considering, aggregating and weighting them in different ways. On the other hand, some works have tried to define specific elasticity metrics. In [5] elasticity is defined as the degree a cloud layer autonomously adapts capacity to workload over time. This work presents a systematic literature review of definitions and metrics for elasticity, but this is not the scope of our research. Nevertheless, some significant examples are introduced in the following paragraphs because they were proposed with a "keep-it-simple" philosophy similar to the one considered in our research.

In [6], elasticity means that more capacity can be added to a running service by deploying new instances of each component and shifting load to them. To quantify this performance figure, in this work an elastic speedup has been defined: a service offering good elasticity should show a performance improvement when new capacity is added, with a short or non-existent period of disruption while the service is reconfiguring itself to use this new capacity.

In [7], the traditional definition of elasticity used in physics is used to represent elasticity in cloud computing and to quantify how fast and efficiently the computing resources are varied in response to users' demand. In this work the stress is defined taking into account the accuracy while the strain is defined considering the speed of the scaling.

In [8] the elasticity is considered an economic aspect of a cloud service besides cost. Elasticity is defined as the capability of both adding and removing resources rapidly in a fine-grain manner. In other words, an elastic cloud service concerns both growth and reduction of workload, and particularly emphasizes the speed of response to changed workload. Due to this essential relationship with the speed, the metrics proposed to quantify elasticity are the Provision (or Deployment) Time and the Boot Time (these two times as components of the Total Acquisition Time) as well as the Suspend Time and the Delete Time (as components of the Total Release Time). Therefore, elasticity is quantified in time units, from 0 to ∞.

In [9] authors claim that cloud users need to know whether a reduced load leads to a reduced bill. They propose to measure elasticity by running a varying workload and comparing the resulting price with the price for the full load. In this case elasticity is measured in monetary units, going from 0 to ∞.

In [10] elasticity is defined as the degree to which a system is able to adapt to workload changes by provisioning and de-provisioning resources in an autonomic manner, such that at each time the available resources match the current demand as closely as possible. Again accuracy and time are considered. Being θ the average time to switch from a system configuration to another and μ be the average percentage of under-provisioned resources during the scaling process, the elasticity (El) is defined as:

$$El = \frac{1}{\theta \cdot \mu}. \tag{1}$$

Elasticity is in this case a metric measured in *time units*$^{-1}$ from 0 to ∞. This definition has been also used in [11] to evaluate and to benchmark cloud elasticity.

In [12] an elasticity metric is supposed to answer these two questions: how often does the system violate its requirements? and once these requirements are violated, how long does it take before the system recovers to a state in which requirements are met again? In this work two metrics are defined to answer these questions, the number of SLO (Service Level Objectives) violations per time unit (from 0 to ∞) and the Mean Time To Quality Repair or MTTQR (in time units, from 0 to ∞).

Finally, in [13] a set of time-related and accuracy-related metrics is proposed to characterize the elasticity of a self-adaptive platform in an IaaS context. Specifically, an aggregated elastic speedup is defined based on accuracy, timeshare and jitter.

As it can be seen, although elasticity definitions and models usually consider that scalability, accuracy, time and cost are its essential components, the elasticity metrics previously defined focus almost completely on time or cost aspects, and very rarely, on accuracy aspects. This implies that a service is considered elastic if it can adapt to the user's needs in a fast and cheap manner. But, what happens if the final configuration of the service is not exactly what the user was asking for? This fact is not taken into account by most of current elasticity metrics explicitly. If someone needs information about this accuracy, it has to be considered independently, as another performance metric. The same happens with scalability. As a result, current cloud users and providers need to handle complex sets of performance metrics when they are interested in different aspects regarding elasticity. This is one of the reasons why this performance figure is rarely quantified, evaluated, analysed, used as Service Level Objective (SLO), monitored and therefore, improved or optimized. As this