Automated Control for SLA-Aware Elastic Clouds

Sara Bouchenak

University of Grenoble – INRIA, Grenoble, France
Currently visiting Universidad Politécnica de Madrid, Spain

Sara. Bouchenak@inria.fr

ABSTRACT

Although Cloud Computing provides a means to support remote, on-demand access to a set of computing resources, its ad-hoc management for quality-of-service and SLA poses significant challenges to the performance, availability and economical costs of the cloud. This paper discusses these issues and presents early ideas to handle them. First, it introduces the *SLAaaS* model (*SLA awa*re *S*ervice) that enriches the general paradigm of Cloud Computing. SLAaaS enables a systematic and transparent integration of service levels and SLA to the cloud. It is orthogonal to IaaS, PaaS and SaaS and may apply to any of them. Furthermore, the paper discusses autonomic SLA management in the cloud and presents early ideas to tackle it.

Categories and Subject Descriptors

C.2.4 [**Distributed systems**]: Client/server; Distributed applications; C.4 [**Performance of systems**]: Modeling techniques; Reliability, availability, and serviceability; I.2.8 [**Problem Solving, Control Methods, and Search**]: Control theory

General Terms

Performance, Reliability

Keywords

Cloud computing, SLA, Performance, Availability, Control, Autonomic elastic clouds, Distributed systems

1. INTRODUCTION

Cloud Computing is a paradigm for enabling remote, on-demand access to a set of configurable computing resources. A Cloud may stand at different levels of the hardware and software stack where: (i) an *Infrastructure as a Service (IaaS)* cloud enables access to hardware resources such as servers and storage devices, (ii) a *Platform as a Service (PaaS)* cloud allows the access to software resources such as operating systems and software development environment, and (iii) a *Software as a Service (SaaS)* cloud is an alternative to classical software applications running locally on personal computers which are, instead, provided remotely by the cloud (e.g. messaging services, document editing services, etc.). A large number of Cloud Computing environments are proposed:

- IaaS clouds such as Amazon S3 and AT&T Synaptic storage services [1][5], Amazon SimpleDB and Microsoft SQL Azure database services [2][14], or Amazon EC2 and AT&T Synaptic Compute computing services (i.e. servers) [3][5];
- PaaS clouds such as Microsoft Azure and Google AppEngine software developments environments [12][7];
- SaaS clouds such as document editing and communication services provided by Microsoft BPOS, Google Apps or HP Cloud Assure [13][8][9].

In this context where multiple clouds provide similar services (e.g. several IaaS clouds provide similar services to access computing resources), it is not easy for a consumer to compare the proposed cloud services and choose the most appropriate one for his needs. We believe that a differentiating element between Cloud Computing solutions will be the quality-of-service (QoS) and the service level agreement (SLA) guaranties provided by the cloud. Existing commercial cloud solutions include some kind of SLA, but express it with vague terms such as "small vs. large instances". Very few QoS aspects are considered in the cloud, for instance, no guaranties regarding performance are provided. Furthermore, the cloud does not automatically handle dynamic variations of cloud usage. Initiatives such as Amazon Auto Scaling help the customer to adapt the size of his cloud [4]. However, significant efforts from the customer are required for cloud capacity planning; this goes against one of the main motivations of Cloud Computing that is the ease of use of services by cloud customers. In summary, Cloud Computing faces the following open issues:

- Need of integration of QoS and SLA requirements with the cloud.
- Automated dynamic elasticity of the cloud for SLA management.

In order to address these issues, we call for: (i) a definition of a new cloud model that integrates SLA as part of the cloud, and (ii) an automated cloud control for building SLA-aware dynamic elastic clouds.

2. SLAaaS CLOUD MODEL

We call for a systematic and transparent integration of quality-of-service and service level agreement to the cloud. Quality-of-service (QoS) in the cloud may refer to several aspects such as performance (e.g. service response time, throughput), or availability (e.g. service abandon rate). Service level agreement (SLA) in the cloud is a contract between a cloud provider and a cloud customer. It specifies the levels of service that the cloud should provide to the customer in terms of objectives to attain for different QoS aspects. Another desirable objective is the reduction of cost of the cloud, i.e. its impact on the energy footprint and the economical costs.

FeBid'10, April 13, 2010, Paris, France.
Copyright © 2010 ACM 978-1-4503-0077-3/10/04 ...$10.00.

We introduce a new model, the *SLA* *a*w*a*re *S*ervice (SLAaaS) model, to enrich the general paradigm of Cloud Computing. SLAaaS is orthogonal to IaaS, PaaS and SaaS clouds and may apply to any of them. With SLAaaS, a cloud clearly exhibits its service levels and the proposed SLA. This enables a consumer who looks for a cloud service to transparently compare service levels of different cloud solutions before choosing the one that is best suited for his needs.

3. SLA-AWARE ELASTIC CLOUDS

In order to guarantee the SLA of SLAaaS clouds, automated control for dynamic cloud elasticity should be provided. This aims to meet quality-of-service requirements such as performance and availability while minimizing cloud cost (i.e. energetic impact and economical costs). In the following, we identify and discuss three complementary research directions to provide SLA-aware dynamic elastic clouds.

Online observation and monitoring of the cloud. This aims to automatically capture variations in cloud usage and workload, to detect SLA violation and to trigger cloud reconfiguration when necessary. QoS measurements may apply at different levels to provide low-level metrics for IaaS clouds or higher application-level metrics for SaaS clouds. The main issue here consists in defining scalable, accurate and non-intrusive distributed algorithms for cloud monitoring.

Modeling the cloud. A cloud has a dynamic behavior with varying and nonlinear cloud service workloads. This has a direct impact on cloud QoS. A cloud is also characterized by its actual configuration, i.e. its size (number of cloud services), its location (machines hosting cloud services), and its service parameters (configuration parameters of individual cloud services). Obviously, the cloud configuration has an impact on both cloud QoS and cloud cost. Cloud modeling aims to render the impact of cloud workload and configuration on the QoS and cost of the cloud. The challenge here is to define a model that is accurate, capable of rendering the nonlinear variation of cloud workload, and that is easy to use with real world clouds (e.g. via automated and online tuning of model parameters). Control theory modeling techniques can be effectively applied here.

Automated control of the cloud. Cloud elasticity is the ability of the cloud to change its configuration, while dynamic cloud elasticity is the ability of the cloud to be elastic while the service is online. Thus, automated cloud control aims to build a dynamic elastic cloud (i.e. a new cloud configuration) that meets QoS requirements as specified in the SLA while minimizing cloud cost (energy footprint, economical costs). To do so, the use of a cloud model allows to reason about the variations of cloud configuration and workload and their impact on cloud QoS and cost. Mathematical optimization and control theory techniques can be effectively used in this context. First, the definition of an objective function allows to precisely quantify SLA QoS requirements vs. cost of a cloud configuration. The use of mathematical optimization techniques allow to determine, for a cloud workload, an optimal configuration, i.e. a cloud configuration that maximizes the objective function by guaranteeing SLA requirements while minimizing the cost.

The definition of control laws describes how to automatically change the cloud configuration to an optimal configuration with respect to the objective function.

The challenge here is multifold: (i) the definition of scalable and optimal control algorithms for the cloud, (ii) the handling of different and sometimes antagonist QoS requirements for the cloud (e.g. performance vs. availability), (iii) the monitoring of the underlying distributed system tackling scalability and accuracy of the monitored data, and (iv) the proposal of techniques for online cloud reconfiguration such as online service (un-)provisioning (i.e. cloud rescaling), online redeployment (e.g. virtual machine migration in IaaS clouds), and online service's internal parameter reconfiguration (e.g. application server parameters in PaaS clouds).

4. CONCLUSION

This paper biefly describes early ideas for a systematic integration of SLA to the cloud through the definition of the SLAaaS cloud, and research directions for an automated cloud control for building SLA-aware dynamic elastic clouds.

5. REFERENCES

[1] Amazon. *Amazon Simple Storage Service (Amazon S3)*. 2009. http://aws.amazon.com/s3/

[2] Amazon. *Amazon SimpleDB*. 2009. http://aws.amazon.com/simpledb/

[3] Amazon. *Amazon Elastic Compute Cloud (Amazon EC2)*. 2009. http://aws.amazon.com/ec2/

[4] Amazon. *Amazon Auto Scaling*. 2009. http://aws.amazon.com/autoscaling/

[5] AT&T. *AT&T Synaptic Storage as a Service*. 2009. https://www.synaptic.att.com/

[6] S. Bouchenak, N. D. Palma, D. Hagimont, and C. Taton. Autonomic Management of Clustered Applications. In *IEEE International Conference on Cluster Computing (IEEE Cluster 2006)*, Barcelona, Spain, Sept. 2006.

[7] Google. *Google App Engine*. 2009. http://code.google.com/intl/fr/appengine/

[8] Google. *Google Apps*. 2009. http://www.google.com/intl/fr/apps/business/

[9] Hewlett-Packard. *Cloud Assure*. 2009. http://www.hp.com/hpinfo/newsroom/press/2009/090331xa.html

[10] L. Malrait, S. Bouchenak, N. Marchand. Fluid Modeling and Control for Server System Performance and Availability. The *39th Annual IEEE International Conference on Dependable Systems and Networks (IEEE DSN 2009)*, Estoril, Lisbon, Portugal, Jun-Jul 2009.

[11] L. Malrait, S. Bouchenak, N. Marchand. Modeling and Control of Server Systems: Application to Database Systems. *The European Control Conference 2009 (ECC 09)*. Budapest, Hungary, Aug 2009.

[12] Microsoft. *Windows Azure Platform*. 2009. http://www.microsoft.com/windowsazure/

[13] Microsoft. *Microsoft BPOS*. 2009. http://www.microsoft.com/online/business-productivity.mspx

[14] Microsoft. *SQL Azure*. 2009. http://www.microsoft.com/windowsazure/sqlazure/