

Automated Control for SLA-Aware Elastic Clouds

Problem statement & Early ideas

Sara Bouchenak

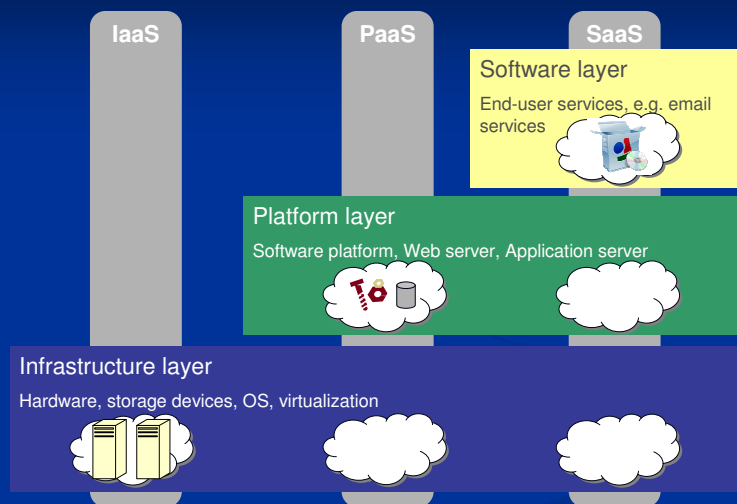
Grenoble University – INRIA

Sara.Bouchenak@inria.fr
<http://sardes.inrialpes.fr/bouchena/>

Cloud Motivations

- The cloud as an alternative to classical:
 - Software applications installed and running locally on personal computers,
 - Hardware environments (e.g. clusters) installed locally for a company, a university, etc.
- Motivations
 - Remote, on-demand access to computing resources
 - Configurable, pay-per-use resources
 - ↔ Ease the use of services by customers
 - ↔ Transparently handle the configuration of services

Cloud Computing Models



Cloud Environments

- Examples
 - IaaS clouds
 - Storage services: Amazon S3, AT&T Synaptic
 - Database services: Amazon SimpleDB, Microsoft SQL Azure
 - Computing servers: Amazon EC2, AT&T Synaptic Compute
 - PaaS clouds
 - Software development environments: Microsoft Azure, Google AppEngine
 - SaaS clouds
 - Document editing and communication services: Microsoft BPOS, Google Apps, HP Cloud Assure

Problem Statement

- Similar services provided by different clouds
 - Not easy for a customer to compare the proposed clouds
- A differentiating element between clouds will be the quality-of-service (QoS) and service level agreement (SLA) guaranties
- Existing commercial cloud solutions include some kind of SLA however:
 - SLA expressed with vague terms (e.g. “small vs. large instances”)
 - Few QoS aspects are considered (e.g. no guaranties regarding performance)
 - No automatic handling of dynamic variations of cloud usage (e.g. significant effort required from the customer for cloud capacity planning)

Problem Statement (2)

- Current state of QoS in the cloud
 - Goes against one of the main motivations of cloud computing, i.e. the ease of use of services by customers
- Cloud computing still fails in providing dynamically elastic clouds
- ↪ Integration of QoS and SLA with the cloud
- ↪ Autonomic control of elastic clouds

Research Directions

- SLAaaS (*SLA-aware Service*)
 - New cloud model, orthogonal to IaaS, PaaS, SaaS
 - SLAaaS clearly exhibits its QoS and SLA
 - Service level objectives: QoS to guarantee
 - Additional objectives: Economical cost and energy footprint to minimize
 - Customers can easily compare different clouds regarding their QoS/SLA

Research Directions – Autonomic Elastic Cloud

- Clouds are complex distributed systems with
 - Variable scale, changing workloads, service composition
- Autonomic elastic clouds from System's and Control's points of view
 - **S:** “Models of computing systems are too simple, do not capture the dynamism of distributed systems and the complexity of clouds”
 - ↪ **C:** Accurate nonlinear modeling techniques (MBM, IEEE DSN 2010)
 - **S:** “Accurate models are too complex to use, require extensive calibration and configuration”
 - ↪ **C:** Automated model calibration to render behavior variation, via systematic and scalable online monitoring techniques (WCO, ACM Middleware 2008, PHZ, EuroSys 2009)
 - **S:** “How to consistently handle different, and sometimes antagonist, QoS requirements (performance vs. availability)”
 - ↪ **C:** Defining optimal control of cloud resources (SLOs and costs)

References

- L. Malrait, S. Bouchenak, N. Marchand. Fluid Modeling and Control for Server System Performance and Availability, IEEE DSN 2010
- P. Padala, Kai-Yuan Hou, Xiaoyun Zhu, Mustafa Uysal, Zhikui Wang, Sharad Singhal, Arif Merchant, Kang G. Shin. Automated Control of Multiple Virtualized Resources, EuroSys 2009
- T. Wood, L. Cherkasova, K. Ozonat, P. Shenoy. Profiling and Modeling Resource Usage of Virtualized Applications, ACM Middleware 2008