2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing

# Towards an SLA-based Service Allocation in Multi-Cloud Environments

Soodeh Farokhi
*Institute of Information Systems*
*Vienna University of Technology, Austria*
*soodeh.farokhi@tuwien.ac.at*

***Abstract*—Cloud computing popularity is growing rapidly and consequently the number of companies offering their services in the form of Software-as-a-Service (SaaS) or Infrastructure-as-a-Service (IaaS) is increasing. The diversity and usage benefits of the IaaS offers are encouraging SaaS providers to lease resources from the Cloud instead of operating their own data centers. This helps them to get rid of the maintenance overheads and better satisfy their customers which are more demanding in terms of service requirements nowadays. Such evolutionary tendency is leading to the emergence of new ways of service provisioning in which relying on infrastructure services of a single Cloud provider is not sufficient. Namely, the need of using Cloud services from multiple Clouds with various quality attributes and pricing models has been raised recently. Although service allocation based on Service Level Agreement (SLA) has been well investigated in Cloud computing so far, the new upcoming issues regarding to utilize multiple Clouds has led to new challenges. This paper looks at the service selection and allocation in a Multi-Cloud, as a delivery model of multiple Clouds, from the perspective of SaaS provider. The proposed framework assists SaaS providers to find suitable infrastructure services which best satisfy their requirements while handling SLA issues. We present an overview of the complete system and discus how the services are selected and the corresponding SLAs are monitored to detect the SLA violations.**

***Keywords*-Cloud Computing, Multi-Cloud, Service Level Agreement (SLA), Service Selection, Service Allocation, Software-as-a-Service (SaaS), Infrastructure-as-a-Service (IaaS), InterCloud-SLA.**

## I. Introduction

A natural evolution in Cloud computing is happening by using different services from multiple Clouds in order to have a wider range of choices with various cost and quality of services (QoSs). Improving the QoS, while optimizing service cost; the ability to migrate among several providers; avoiding vendor lock-in; and the need of particular Cloud services which are not provided elsewhere are some of the reasons for using services from multiple Clouds. In general, two types of delivery models exist in multiple Clouds: Federated Cloud and Multi-Cloud [17], which differ in the degree of collaborations between the involved Clouds and the way that the user interacts with them. In the Federated model, an agreement between the involved Cloud providers, transparent from the user, is required. While, in the Multi-Cloud model, as the focus of this work, there is no need for such agreement. Furthermore, in the Multi-Cloud model, users are aware of multiple Clouds, and usually a third party, named as a Multi-Cloud middleware in this work, is responsible to deal with Cloud provider API variations.

The advantages of utilizing Cloud infrastructure services and especially the diversity of pricing models and QoSs in a Multi-Cloud are encouraging software providers to exploit this fertile environment. Meanwhile, SaaS providers are looking into solutions that minimize the overall infrastructure leasing cost without adversely affecting their customers [20]. To achieve this goal, it is essential to have a clear definition of SaaS provider requirements, so in this context, Service Level Agreement (SLA) serves as a foundation for the expected functional and quality level of the service between the involving parties [14] (IaaS and SaaS providers in our case). Considering SLA in Cloud service allocation makes Cloud computing a valid alternative model to the private data centers regarding to user QoS requirements such as availability and security [5].

A thorough literature review conducted by the author on service selection and SLA management revealed two main problems: First, although the SLA-based service selection in Cloud computing is well researched and analyzed, an approach in which the SaaS provider profit is the main focus and it utilizes the promising features of a Multi-Cloud environment is missing. Namely, existing works have mostly focused on maximizing the profit of either the customers or the IaaS providers by proposing solutions in a single Cloud without thoroughly investigating SLA issues [3], [6]. Thus, barriers relevant to the SLA management while selecting and allocating services, such as SLA interoperability, SLA validation tracking, SLA violation detection in a Multi-Cloud, from the SaaS provider perspective, has not been explored yet. As a recent investigation of SLA interoperability issues in Multi-Cloud, an IEEE working group called InterCloud Working Group (ICWG)[1], has been established to develop a set of standards for InterCloud interoperability.

Second, the growth of Cloud providers both in the number of players and the variety of offered services forces companies to deal with a not trivial selection problem [7]. Since by using a Multi-Cloud environment, dependent components of a single SaaS application can be distributed in different Cloud data centers, with various QoS attributes, from the SaaS provider perspective, the deployment can be considered as an abstract *Composite Infrastructure Service* with a set of functional and non-functional requirements for each included component. The question remains how to, on the one hand, score and select services for each single component

[1] http://grouper.ieee.org/groups/2302/

978-1-4799-2784-5/14 $31.00 © 2014 IEEE
DOI 10.1109/CCGrid.2014.62

and on the other hand, optimize this allocation to satisfy the requirements of the composite service. Furthermore, these concerned QoS parameters can be conflicting or have various importance degrees for the SaaS provider.

Our work, as a recently started PhD thesis, ultimately aims to provide solutions to the above mentioned problems by proposing a framework for service allocation in a Multi-Cloud environment while taking into consideration the SLA management issues.

The remainder of this paper is organized as follows. Section II briefly presents the related work. Section III describes the overview of the proposed framework. In Sections IV and V primarily results and evaluation plan are discussed. Finally, Section VI concludes the paper.

## II. State of the Art

In recent years, extensive research has been conducted in the area of service selection and SLA management in Cloud computing environments. However, for the service selection in a Multi-Cloud environment, a methodology is needed to compare Cloud services based on the various criteria such as cost and QoS parameters for different user profiles [17]. In addition, due to the SLA heterogeneity in this environment, SLA management from whether customer or provider perspectives is a challenging task.

Most of the approaches that focus on the SLA-based service selection and allocation in the Cloud are trying to maximize the customer profit [6], [8] or IaaS provider profit [15]. The work in [20] is one of the first attempts dealing with resource allocation from the SaaS provider perspective. The authors propose an allocation strategy for SaaS providers to maximize their profits and customer satisfaction levels when deploying their applications on the Cloud infrastructure services. However, from the SLA perspective, they only consider response time and service initiation time. In addition, the evaluation of this work is performed in a single Cloud with one virtual machine (VM) request per service.

Similar work [19] has investigated service allocation in the Cloud by providing an SLA-driven resource allocation scheme that selects a proper data center among globally distributed centers operated by a single provider. In contrast, we support composite Multi-Cloud services where the SLA can include several parameters such as availability, latency, reputation, throughput and cost.

An extensive evaluation of existing approaches dealing with SLA in Cloud computing has been done recently in [14]. Among research projects introduced in this report, the Contrail project [5], [7] has similar goals to our proposed work regarding to the SLA management for composite services in a multiple-provider environment with different resource types. However, our focus is on a different multiple Cloud delivery type, Multi-Cloud model, while this project works on the Federated-Cloud model. Moreover, aside the different needs for SLA interoperability in these two models, the main goal of the Contrail project is to allow Cloud providers to seamlessly integrate resources from other Clouds with their own infrastructures, and break the current customer lock-in situation by allowing live application migration from one Cloud to another. While, our goal is providing a framework, as a middleware for the SaaS provider, in order to minimize the infrastructure leasing cost and SLA violation rate as well as maximizing the satisfaction level by utilizing the Cloud infrastructure services of a Multi-Cloud environment.

## III. Multi-Cloud Service Allocation Framework

As depicted in Figure 1, the proposed framework lies between the SaaS and IaaS provider layers and manipulates the Multi-Cloud service selection and SLA management. As an accepted third party of Multi-Cloud delivery model, we assume all transactions between the proposed framework and the IaaS providers are done through the Multi-Cloud middleware that communicates with the APIs of all involved Cloud providers.

The *SLA Construction Engine* and *Service Selection Engine* handle the issues surrounding the SLA formation and service selection by communicating with the *SLA Repository* and *IaaS provider Profiles Repository* components. While, other components cover the SLA validation tracking, violation detection and enforcement. These will be done by monitoring the allocated and running services and applying certain strategies in order to detect SLA violation and then react towards them which is the responsibility of *SLA Validation and Enforcement Engine*.

The whole service allocation process will be realized in the framework through three main phases: (1) *SLA Construction*, (2) *Service Selection*, and (3) *SLA Monitoring and Violation Detection*. The first phase forms SLAs named InterCloud-SLAs[2], which are provider-independents. These SLAs include SaaS provider's QoS requirements for the deployment of the application on the Cloud. Furthermore, the functional requirements are expressed with the Open Virtualization Format (OVF) standard [9]. The second phase uses a selection algorithm to score services based on the user satisfaction level for each service. The principle of prospect theory [13] is used in the selection algorithm to model the user satisfaction as a function of service QoS parameters and the importance of each parameter for the user. While the first two phases are at design time, the last phase deals with the monitoring of the allocated services in order to detect, and in some cases avoid, the SLA violation at runtime. Since, the requested services of a SaaS provider can be selected from more than one Cloud infrastructure provider in a Multi-Cloud environment, the service monitoring and SLA

[2] Inspired by a sub-group of the aforementioned IEEE intercloud working group, named as InterCloud-SLA.

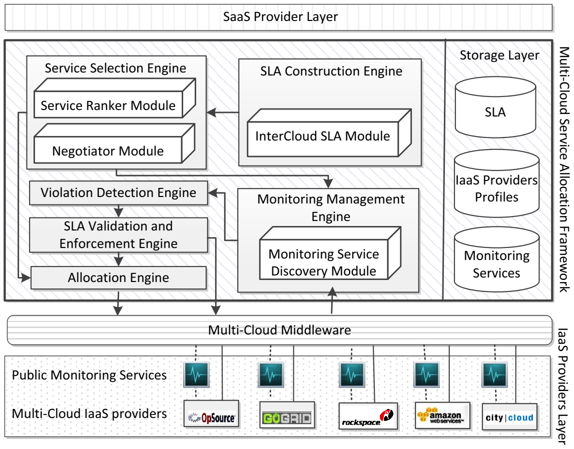

Figure 1: Multi-Cloud service allocation framework.

detection encompass tedious challenges which are going to be tackled in this phase.

***SLA Construction Phase*** As the input of this phase, the SaaS provider submits its Cloud infrastructure requirements to the framework as a single XML file. These requirements contain two parts: one includes the requirements for each single infrastructure service (Cloud VM or storage), and the other one contains the requirements of the composite infrastructure service. The data related to these two parts is extracted from the given XML file and transformed into a set of SLAs by using Model Driven Architecture (MDA) [16] principles and the specification of OVF standard for the functional requirements. The main purpose of constructing such SLAs is addressing the SLA interoperability issue in a Multi-Cloud. Based on the MDA models, InterCloud-SLA can be considered as a Platform Independent Model (PIM) while each IaaS provider's SLA can be modeled as a Platform Specified Model (PMS). The goal of this phase is first modeling the PIMs and then automatically transforming them to the corresponding PSMs of selected infrastructure services. Moreover, some sort of reasoning abilities are required to be applied in this phase in order to break down the SaaS requirements in a way that they can be mapped to the available service offers, or to the combinations of offers in order to provide new value-added services for the user requirements.

***Service Selection Phase*** The InterCloud-SLAs and the specification of IaaS providers' offers are two inputs of this phase. The service selection algorithm used in this phase, first finds the best set of services which satisfy the requirements of each involved service. Afterward, it considers the whole request as a composite service and tries to choose an optimum combination of services by utilizing multiple Cloud providers. Considering the latency and data traffic issues among included selected services of the composite service at runtime in a Multi-Cloud is one of the key challenges of this phase. We believe the relation between service QoS parameters and the user satisfaction can be modeled effectively and precisely by using the principles of prospect theory.

***SLA Monitoring and Violation Detection Phase*** As emphasized in [2]: An SLA cannot guarantee that the service is delivered as it has been described, similar to the case that a car guarantee cannot claim that your car will never break down. ”In particular, an SLA cannot make a good service out of a bad one. However, it can mitigate the risk of choosing a bad service.” Sticking to this goal for using SLA, this phase is responsible for handling the SLA management tasks at runtime includes SLA monitoring, SLA validation tracking, SLA violation detection, and SLA enforcement. Possible monitoring strategies include developing APIs to provide a unified monitoring on Cloud vendors or enabling Trusted Third Parties (TTP) to undertake the monitoring responsibilities [14].

SLA validation tracking can be done by utilizing an Abstract Behavioral Specification (ABS) language [11]. ABS is a high-level, executable programming languages, which is used to support full code generation and (timed) validation of models [1]. In our work, it will be used for the validation of SLAs in a Multi-Cloud environment.

SLA violation detection or detection of the future violations will be done by reasoning on the gathered information of the monitoring services. Some strategies such as service migration and defining a penalty model to influence the provider reputation can be applied at this phase. Among possible solutions for the SLA violation detection, we are investigating the application of modeling the problem as a Root Cause Analysis (RCA) problem in Bayesian Networks [18].

## IV. Preliminary Results

So far, we have designed and implemented the involved modules of the *Service Selection Phase* in [10]. The cornerstone of this work is a novel service selection algorithm that works based on prospect theory in order to compute the user satisfaction score for a certain service. The proposed algorithm scores the infrastructure services (Cloud VM or storage) based on the user satisfaction degree by considering the service QoSs and SLA parameters. As prospect theory is an alternative decision making model for utility theory and is said to be more realistic in calculating the user satisfaction [13], we have evaluated the proposed algorithm with a state-of-the-art, utility-based algorithm [12]. This comparison was done based on the implementation of both algorithms and evaluating the selected services in a simulated environment. This simulation was enriched by realistic data from the commercial Cloud IaaS providers. The result showed that our approach selects a set of services that more effectively satisfy constructed InterCloud-SLAs.

## V. Evaluation Plan

The efficiency of the InterCloud-SLAs, as the outputs of *SLA Construction Phase*, is evaluated by their ability to support various Cloud IaaS providers' SLAs in the form of PSM, from the MDA perspective. Furthermore, the automation level of model transformation from the PIM (InterCloud-SLA) to the corresponding PSM (provider SLA) will be another evaluation factor for this phase.

While the second phase has been already partially examined in [10], to evaluate the last phase which is *SLA monitoring and violation detection*, we will first use the CloudSim toolkit [4] to model multiple infrastructure services includes VM and storage offers, each with different pricing models and QoS parameters. Afterwards, we will assay the accuracy of monitoring service discovery algorithm by the number of successful matching. SLA validation will be evaluated by using the ABS language to support timed validation of SLAs corresponding to the selected IaaS providers. For the further steps, we will model the SLA violation detection process as a RCA problem and solve it by using the existing machine learning approaches such as Bayesian Network. We will evaluate the effectiveness of this modeling by measuring the rate of SLA violation at runtime.

## VI. Conclusion

The diversity of services in a multiple Clouds environment is encouraging more SaaS providers to move towards using the infrastructure services provided by the Cloud providers instead of running their own data centers. However, the lack of an efficient service allocation and SLA management approach that maximizes SaaS providers' benefits in a Multi-Cloud environment, as a delivery model of multiple Clouds, impedes this evolutionary process. To tackle these barriers, in this work, we proposed a Multi-Cloud service allocation framework which contains three main phases *SLA Construction*, *Service Selection* and *SLA Monitoring and Violation Detection*. This paper is a preliminary schema of the proposed framework, so there are still many challenges which are needed to be covered during the complete definition and implementation of the system.

## Acknowledgment

This work is partly financed by the "Adaptive Distributed Systems" doctoral program, Vienna University of Technology. Soodeh Farokhi is a first year PhD candidate in Computer Science and is jointly supervised by Dr. Ivona Brandic and Prof. Helmut Veith.

## References

[1] E. Albert, F. S. de Boer, R. Hähnle, E. B. Johnsen, R. Schlatte, S. L. T. Tarifa, and P. Y. Wong, "Formal modeling and analysis of resource management for cloud architectures: an industrial case study using Real-Time ABS," *Service Oriented Computing and Applications*, pp. 1–17.

[2] P. Allen, *Service orientation: winning strategies and best practices*. Cambridge University Press, 2006.

[3] P. Bellavista, A. Corradi, L. Foschini, and A. Pernafini, "Automated Provisioning of SaaS Applications over IaaS-Based Cloud Systems," in *Advances in Service-Oriented and Cloud Computing*. Springer, 2013, pp. 94–105.

[4] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. De Rose, and R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and Experience*, vol. 41, no. 1, pp. 23–50, 2011.

[5] R. G. Cascella, L. Blasi, Y. Jegou, M. Coppola, and C. Morin, "Contrail: Distributed Application Deployment under SLA in Federated Heterogeneous Clouds," in *The Future Internet*. Springer, 2013, pp. 91–103.

[6] K. Clark, M. Warnier, and F. M. Brazier, "Automated Non-repudiable Cloud Resource Allocation," in *Cloud Computing and Services Science*. Springer, 2013, pp. 168–182.

[7] C. Consortium, "Overview of the Contrail system, components and usage," Contrail consortium, Tech. Rep., 01 2014.

[8] V. Dastjerdi, "QoS-aware and semantic-based service coordination for multi-Cloud environments," Ph.D. dissertation, University of Melbourne, march 2013.

[9] Distributed Management Task Force, "Open Virtualization Format Specification," Tech. Rep., 2010. [Online]. Available: http://www.dmtf.org/sites/default/files/standards/documents/DSP0243_1.1.0.pdf

[10] S. Farokhi, F. Jrad, I. Brandic, and A. Streit, "HS4MC: Hierarchical SLA-based Service Selection in Multi-Cloud Environments," in *4th International Conference on Cloud Computing and Services Science Special Session on Multi-Cloud*. CLOSER, 2014.

[11] E. B. Johnsen, R. Hähnle, J. Schäfer, R. Schlatte, and M. Steffen, "ABS: A core language for abstract behavioral specification," in *Formal Methods for Components and Objects*. Springer, 2012, pp. 142–164.

[12] F. Jrad, J. Tao, R. Knapper, C. M. Flath, and A. Streit, "A utility-based approach for customised cloud service selection," *Int. J. Computational Science and Engineering*, 2013.

[13] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica: Journal of the Econometric Society*, pp. 263–291, 1979.

[14] D. Kyriazis, "Cloud Computing Service Level Agreements, Exploitation of Research Results," European Commission Directorate General Communications Networks Content and Technology Unit, Tech. Rep., 05 2013.

[15] Y. C. Lee, C. Wang, A. Y. Zomaya, and B. B. Zhou, "Profit-driven service request scheduling in clouds," in *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*. IEEE Computer Society, 2010, pp. 15–24.

[16] Object Management Group (OMG). (2014, March) MDA Specifications. [Online]. Available: http://www.omg.org/mda/specs.htm

[17] D. Petcu, "Multi-Cloud: expectations and current approaches," in *Proceedings of the 2013 international workshop on Multi-cloud applications and federated clouds*. ACM, 2013, pp. 1–6.

[18] O. Pourret, P. Naïm, and B. Marcot, *Bayesian networks: a practical guide to applications*. Wiley. com, 2008, vol. 73.

[19] S. Son and S. C. Jun, "Negotiation-based Flexible SLA Establishment with SLA-driven Resource Allocation in Cloud Computing," in *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on*. IEEE, 2013, pp. 168–171.

[20] L. Wu, S. K. Garg, and R. Buyya, "SLA-based resource allocation for software as a service provider (SaaS) in cloud computing environments," in *Cluster, Cloud and Grid Computing (CCGrid), 2011 11th IEEE/ACM International Symposium on*. IEEE, 2011, pp. 195–204.