Future Generation Computer Systems 55 (2016) 186–199

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Capacity-driven utility model for service level agreement negotiation of cloud services

Nadia Ranaldo, Eugenio Zimeo *

Department of Engineering, University of Sannio, Italy

HIGHLIGHTS

- We propose a capacity-aware utility model to support negotiation of cloud services.
- The utility function takes into consideration the available resources dynamically.
- The approach improves the provider utility and reduces SLA violations.

ARTICLE INFO

Article history:
Received 15 May 2014
Received in revised form
18 January 2015
Accepted 9 March 2015
Available online 7 April 2015

Keywords:
Cloud computing
QoS management
SLA
Negotiation
Capacity planning

ABSTRACT

Dynamic customers' requirements and providers' resources availability in the Cloud market make it inadequate static approaches to guarantee Quality of Service (QoS) levels and to define pricing. In this context, negotiation guided by dynamic information is a viable way to achieve high satisfaction levels for both contract parties. We propose to exploit capacity planning to support bilateral negotiation processes with the aim of optimizing the utility for service providers, by avoiding contracts that could incur in Service Level Agreements (SLAs) violations, keeping, at the same time, competitive prices. The proposed technique exploits a non-additive utility function defined in the region of acceptable SLA proposals, taking into account desired QoS and expected resources availability, costs and penalties. The experimental analysis shows the benefit of the proposed dynamic approach with respect to static ones in a scenario characterized by a set of customers and differentiated classes of applications provided by a cloud environment.

1. Introduction

Service Level Agreements (SLAs) [1] represent key elements to achieve full success in Cloud computing, since they represent the desired guarantees between service providers and customers. SLAs allow to formally describe the offered functions, the QoS levels the provider promises to meet, the responsibilities [2] of both the contract parties, and the penalties applied in case QoS levels are not satisfied.

Platform as Service (PaaS) providers (e.g. Google App Engine and Force.com), often offer a pool of differentiated services with pre-fixed prices, related to the complexity of the deployed applications, measured through metrics such as the number of applications and database objects. For these services, SLAs are currently used to define only the granted service availability (uptime) level and a credit-based penalty system in case of violation. They do not offer, yet, the possibility to define custom agreements that could better satisfy both customers and providers.

Coarse-grained and static QoS guarantees are no longer satisfactory in a market characterized by continuously changing conditions. They, in fact, require providers to quickly react in order to maintain high levels of competitiveness and customer satisfaction (birth of new high competitive providers, customers' demand of cloud services for new business fields, fluctuations of electrical power price, optimal data center resources exploitation).

In this dynamic scenario, negotiation of fine-grained SLAs could be a viable approach for service providers to be competitive and to reach more profitable agreements for both customers and providers [3].

The level of flexibility of the negotiation process depends on the underlying protocol. It could be (1) *unilateral*, if a party (typically the provider) proposes a SLA and the other one can only decide to accept or reject it, or (2) *bilateral*, if both the parties have an active role in proposing and defining SLAs. The latter allows to resolve

* Correspondence to: Department of Engineering, University of Sannio, via Traiano, Benevento, 82100, Italy. Tel.: +39 0824 305538; fax: +39 0824 30552.
E-mail addresses: ranaldo@unisannio.it (N. Ranaldo), zimeo@unisannio.it (E. Zimeo).

http://dx.doi.org/10.1016/j.future.2015.03.007
0167-739X/

conflicts deriving by different and continuously changing goals, policies and preferences of customers and providers through dialog between them.

In many bilateral negotiation strategies, each negotiation actor adopts a decision model based on a utility function, which represents the (perceived) satisfaction level associated to a SLA proposal. In particular, given n negotiable SLA parameters, the utility function assigns a utility value to each point in the corresponding n-dimensional space of such parameters. The region in such space in which the utility value is considered acceptable during the negotiation process is called *acceptable region*. Each point in this region represents a SLA proposal and has a utility value between a minimum and a maximum.

Since customers and providers adopt different utility functions that are not known to the counter-parts, an agreement is possible only if the intersection between the two acceptable regions, called *negotiation space*, is not empty. In this case, an agreement is a point in the negotiation space where the utility assumes a satisfactory value for both customer and provider.

An agreement is reached through a process, typically based on time. For example, time-based decision functions [4] allow to make time-dependent concessions with respect to an initial utility value (e.g. the maximum one) with the aim to reach the agreement within a prefixed negotiation time. In particular, when a SLA proposal is received from a negotiation party, the parameters values are verified against the acceptable region and, if they are admissible, the related utility value is computed. On the basis of such evaluations and of elapsed time, the strategy makes decisions about the acceptance or rejection of the proposal, the counter-proposal generation or negotiation termination.

In the literature, typically, the decision models are based on multi- and independent-attribute utility functions that are *additive* with respect to negotiation parameters, that is, the utility can be evaluated considering one parameter at a time, and the total utility can be computed by adding (linear combination) the utility contributions derived from the value of each negotiable parameter [5]. With this approach, a SLA proposal is acceptable if each negotiable parameter value is within the corresponding interval of acceptable values.

In a more realistic Cloud market, some negotiable parameters, such as price and QoS levels, cannot be considered additive independent: the service price depends on resources cost, that, in turn, depends on the agreed QoS terms. Moreover, utility should take into account strategic business policies and dynamic information about the negotiation context, such as market trend, actual customers' requirements and providers' resources availability and performance. In fact, before a SLA is signed, the provider has to check whether the requested set of resources will be available when desired, to avoid future SLA violations. Moreover, an offer with the same QoS level and price could be accepted (refused) on the basis of different conditions: sustainable (not sustainable) service usage conditions (e.g. the forecasted daily load peak) and a high (low) competitive market phase, also in case of potential economic loss.

In this paper, we focus on bilateral SLA negotiation of PaaS services for hosting multi-tier Web applications in a scenario where the number of users is variable and the workload is not stationary but, typically, exhibits peaks and dips with daily, weekly or also seasonal cycles [6,7].

In order to meet QoS terms, the provider allocates appropriate resources to each tier of the application architecture. Currently, we adopt replication only at the application server tier, while a Web server is used as a load balancer and a single database server is shared. Thanks to virtualization technologies, replication is dynamically managed by a resource management system which handles a set of independent and homogeneous virtual machines (the overall Cloud provider capacity).

The virtual machines, allocated on a set of hardware resources of the provider data center, are exploited to host various instances of the application server. The number of virtual machines, allocated to the application server tier of each signed SLA, changes dynamically during the day by means of a predictive resource allocation mechanism. This mechanism aims to define the best resource allocation plan able to maximize the profit and to avoid QoS violations under a daily fluctuating workload.

The proposed utility model, which dynamically defines the acceptable regions on the basis of available capacity, is used at provider-side to guide negotiation strategies. To this aim, the adopted utility function is non-additive to represent the overall provider economic profit deriving by a new contract, net of cost of assigned capacity, penalty payment in case of QoS guarantees violations and eventual variation in profits of already signed SLAs.

Utility is a function of two negotiable parameters, which are the contract price (una-tantum payment) and the maximum response time that can be perceived by the end-user without incurring in a penalty, and other non-negotiable parameters (constraints and pre-conditions). These constraints and preconditions are defined by both customer (contract duration and starting day, application component size, forecasted daily workload plan) and provider (service availability and penalty).

Price is a function of capacity cost and market conditions. Capacity cost depends on the product *virtual machines x daily time slots* assigned to a SLA, whereas market conditions (monopoly vs competition) are captured by using two factors that express (a) the probability to choose that provider and (b) the possible profit.

From the considerations above, the proposed utility function is based on effective customer's requirements, specified in the initial negotiation phase (as pre-conditions), and on a capacity planning technique, which suggests the best profitable resources allocation plan for every new SLA by avoiding (or reducing) violations.

To validate the proposal and to show the benefit in predicting utility of new potential contracts, an in-depth experimental analysis has been conducted.

1.1. Main contribution

To the best of our knowledge, this is the first proposal that adopts capacity planning in the first phase of a contract life-cycle to guide bilateral negotiation strategies through the definition of the providers' acceptable region and utility value. By adopting this approach, the provider reduces the risk of incurring in SLA violations since the technique allows to find actual free slots in the global resource allocation plan, which is defined by considering the resources needed to satisfy all the signed SLAs. Our proposal, unlike the traditional ones based on additive and static utility functions, allows the provider to propose, during negotiation processes, offers with competitive prices and feasible performance. Moreover, it maintains the potential violation of QoS terms under fixed tolerable levels and avoids the stipulation of new contracts in case they conduct to unprofitable revenues or customer unsatisfaction.

Our proposal was firstly presented in [8], in which a preliminary experimental validation, based on a simple linear application performance model was adopted to investigate the proposed utility function and the capacity-driven evaluation technique. A more realistic experimental scenario and the comparison of the proposed approach with the traditional one based on additive utility functions have been presented in [9].

This paper extends both the previous ones, giving deeper details about the utility function formalization and the heuristic adopted for its evaluation, and presenting an in-depth experimental analysis to validate the approach. The analysis shows the achievement of high satisfaction levels for both providers and customers: *providers can gain advantages both in the short period*