2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing

Negotiation-based Flexible SLA Establishment with SLA-driven Resource Allocation in Cloud Computing

Seokho Son
School of Information and Communications
Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
shson@gist.ac.kr

Sung Chan Jun
School of Information and Communications
Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
scjun@gist.ac.kr

***Abstract*—As various consumers tend to use personalized Cloud services, Service Level Agreements (SLAs) emerge as a key aspect in Cloud and Utility computing. The objectives of this doctoral research are 1) to support a flexible establishment of SLAs that enhances the utility of SLAs for both providers and consumers, and 2) to manage Cloud resources to prevent SLA violations. Because consumers and providers may be independent bodies, some mechanisms are necessary to resolve different preferences when they establish a SLA. Thus, we designed a Cloud SLA negotiation mechanism for interactive and flexible SLA establishment. The novelty of this SLA negotiation mechanism is that it can support advanced multi-issue negotiation that includes time slot and price negotiations. In addition, to prevent SLA violations, we provided a SLA-driven resource allocation scheme that selects a proper data center among globally distributed centers operated by a provider. Empirical results showed that the proposed SLA negotiation mechanism supports faster agreements and achieves higher utilities. Also, the proposed SLA-driven resource allocation scheme performs better in terms of SLA violations and the provider's profits.**

Keywords-Cloud Computing; SLA Negotiation; Cost Models of Cloud; Cloud Resource Allocation; Distributed Data Centers

I. INTRODUCTION

Cloud computing is an evolving paradigm to provide consumers with a new utility as various computing services (e.g., Software, Infrastructure, and Platform as a Service). In the Cloud market, consumers are varied and thus have personalized budget plans and requirements for service quality. Also, Cloud service providers (CSPs) have different resource capacities and marketing strategies. As various consumers tend to use personalized services, Service Level Agreements (SLAs) emerge as a key aspect in Cloud computing. There are some standards to support SLAs such as Web Service Agreement Specification (WS-Agreement) [1]. However, SLA-driven Cloud computing designed to enhance utility for both consumers and CSPs is not maturely developed at this time.

Therefore, *the objectives* of this doctoral research focused on two aspects of SLA-driven Cloud computing: 1) supporting SLA establishment (SLA-E) that enhances the utility of the agreements for both CSPs and consumers (i.e., negotiation-based SLA-E), and 2) supporting SLA management (SLA-M) to prevent SLA violations (i.e., SLA-driven resource allocation).

As participants in a Cloud may be independent bodies, in order to establish a flexible SLA, some mechanisms must be in place to resolve the different preferences of those entities. A negotiation mechanism is effective in resolving those different preferences. Whereas it is essential for both a consumer and a CSP to reach an agreement on the price of a service, when to use the service, and Cloud Quality of Service (QoS) issues, to date there is little or no negotiation support for Cloud service reservations with respect to concurrent price, time slot, and QoS negotiation. The purpose of this dissertation—to design a negotiation mechanism that facilitates SLA-E—includes: 1) the design of a multi-attribute negotiation mechanism that takes into account concurrently: price, time slot and QoS, 2) tradeoff algorithms that facilitate decision making in a multi-attribute negotiation, and 3) a one-to-many negotiation mechanism to facilitate distributed Cloud resource allocation.

In addition to facilitating SLA-E, it is important for CSPs to manage limited resources to guarantee the SLAs. Existing CSPs have been deploying and operating data centers globally. Because the resource capacity of a data center is limited, distributing the load to global data centers will provide stable services. Although various load-balancing algorithms have been developed, it is important to avoid SLA violations (e.g., response time) when a CSP allocates the load to data centers around the world. Considering load balancing and guaranteed SLA, therefore, this dissertation proposes 4) an SLA-driven Cloud computing to facilitate resource allocation that takes into account the workload and geographical location of distributed data centers.

II. SIGNIFICANCE OF OUR RESEARCH

Buyya et al. [2] addressed the necessity of SLA-driven (oriented) resource allocation to realize Cloud and Utility computing. They present the challenges and architectural elements of SLA-driven resource management. Along with [2], this research aims to enhance SLA-driven Cloud computing. Whereas [2] provide a SLA-driven Cloud framework that incorporates the challenges, it is important to establish a well-adjusted and mutually agreeable SLA before managing Cloud resources. Accordingly, we focused on both SLA-E and SLA-M in Cloud computing (Fig. 1).

978-0-7695-4996-5/13 $26.00 © 2013 IEEE
DOI 10.1109/CCGrid.2013.81

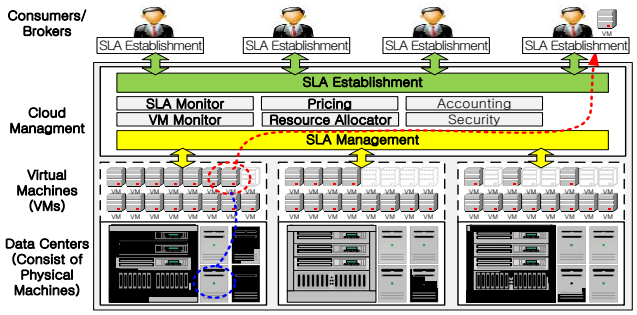

Figure 1. Focus of the doctoral dissertation.

To enhance the utility of SLA-E, we designed a multi-attribute negotiation mechanism that considers price, time slot and Cloud QoS concurrently. Whereas CSPs such as EC2 provide a pre-defined SLA that incorporates fixed price (EC2 also supports auction-based spot price), fixed response time, and some selective performance options, this may restrict diversifying service types and expressing required service level exactly. Thus, SLAs should be variable and flexible to personalize service qualities by budget plans.

Hence, the significance of this research in leveraging such limitations is that with the proposed multi-attribute negotiation, CSPs can support flexible and interactive SLAs. As the proposed mechanism includes time slot negotiation capability, consumers and CSPs can express their temporal preferences in SLAs. Lastly, it can be used as a pricing method that takes into account changing price rates according to market (supply/demand) and resource capability.

Also, it is important to guarantee the established SLAs. As such, we developed an SLA-driven Cloud framework that includes the automated SLA negotiation mechanism and a workload- and location-aware resource allocation (i.e., initial VM placement in a data center). Using the proposed system, a consumer can establish the SLA with respect to service price, time slot, and response time through an automated SLA negotiation; further, a CSP can facilitate load balancing using a pricing strategy. We documented the effectiveness of SLA negotiation and SLA-driven resource allocation in terms of SLA violations and the CSP's profits in where a CSP operates multiple data centers worldwide.

Because we provide a negotiation-based pricing model to Clouds, our research is relevant to 1) the area of economic and utility computing models for Clouds. Also, the SLA-M scheme is included in 2) the topics on scheduling, load balancing and resource management paradigms (both are included in the CCGrid symposium topic areas).

III. Related Work

As this work explores the issue of designing the negotiation-based SLA-E and SLA-driven resource allocations, areas related to this work include: 1) automated negotiation mechanisms and frameworks applied to Grid/Cloud and 2) SLA-driven resource allocation schemes.

1) Automated negotiation in Grid/Cloud computing: There are several automated negotiation mechanisms for Grid/Cloud (see [3] for a survey). Although there are single-issue ([4][5]) and multi-issue negotiation mechanisms [6][7]) for Grid resource negotiation, none of these works considers time slot negotiation. In many existing negotiation mechanisms, a utility function is used to characterize a price utility. The difference between this work and previous researches that consider single [4][5] and multi-issue negotiations without a specific tradeoff algorithm [6][7] is that this work considers a price, time slot, and Cloud QoS issue negotiation concurrently with the design of utility functions and an advanced tradeoff algorithm. Venugopal et al. [8] adopted a protocol for negotiating SLAs based on Rubinstein's alternating offers protocol [9] for the advance reservation of Grid/Cloud resources. Whereas [8] proposed time slot-based resource allocations, the resource allocation focuses on finding a time slot that can be co-allocated; that form of time slot negotiation is not addressed.

For SLA specifications, a meta-negotiation was proposed by Brandic et al. [10] to allow two parties to reach an agreement on what specific negotiation protocols to use before starting the actual negotiation. [11] proposed a declarative rule-based SLA language for describing SLAs generically. Whereas [10][11] do not focus on specifying negotiation strategies or designing utility functions for each negotiation term, [12] proposed a framework for a Web service composition that provides SLA negotiation for QoS constraints. In [12], a utility function-based decision making model is presented. [12] designed a single attribute utility function for linear and monotonic QoS attributes. This function is appropriate for generic attributes (e.g., price). However, here we consider a time slot attribute that is difficult to represent as a linear and monotonic utility function. Also, we designed a trade-off algorithm to enhance negotiation utility and speed.

2) SLA-driven Cloud computing that includes load balancing in global data centers: Sotomayor et al. [13] compared OpenNebula with several well-known virtual infrastructure managers, including Amazon EC2, vSphere, Nimbus, Eucalyptus, and oVirt. The comparison includes resource allocation policies such as static-greedy, round robin, and resource placement considering average CPU load. While the placement focused on selecting a physical machine at a data center, they did not focus on placement to select a proper data center among global data centers. With data centers, we need to consider SLA violations (e.g., response time) because of the network speed. Moreover, existing CSPs such as EC2 do not employ sophisticated VM placement for global data centers, and users themselves manually select a data center at which to place their VMs.

Buyya et al. investigated energy-aware resource provisioning and allocation algorithms to improve the energy efficiency of the data center without violating the negotiated SLA [14]. Whereas [14] provides a research direction for resource allocation in Cloud, [14] does not consider a CSP that operates distributed data centers to balance the resource load and response time by geographical distance, and does not provide a specific SLA negotiation. Le et al. [15] considered load placement policies to manage center temperatures among CSPs operating multiple data centers worldwide. However, While [15] proposes dynamic load distribution policies, there is no focus on the SLA guarantees.

IV. RESEARCH ACCOMPLISHMENTS

A. SLA-driven Cloud computing

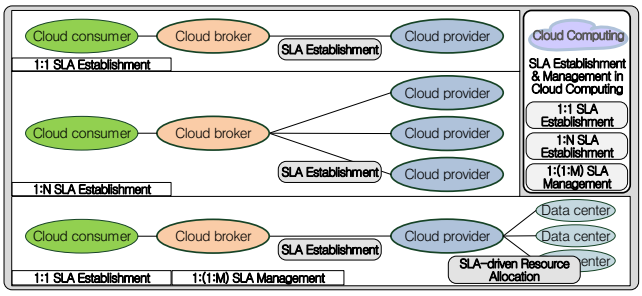

Figure 2. Design models for SLA-driven Cloud computing.

Fig. 2 shows design models included in this research. Each model consists of consumer, broker, and CSP. This includes SLA-E between a consumer and CSP (1:1 SLA-E), SLA-E for a consumer among multiple CSPs (1:N SLA-E), and a 1:1 SLA-E model with resource allocation to multiple data centers (SLA based 1:1:M). In this paper, we introduce the methodology and the major research accomplishments published in [16][17] for 1) Negotiation-based SLA-E and 2) SLA-driven resource allocation in global data centers.

B. Negotiation-based SLA-E

A negotiation mechanism consists of a protocol, strategy, and utility functions. The protocol is a set of communication rules for negotiations. The negotiation mechanism in this work follows Rubinstein's alternating offers protocol [9], which permits agents to make counter-offers to their opponents in alternate rounds. Both agents generate counter-offers and evaluate their opponent's offer. Counter-proposals are generated by the strategy (concession and tradeoff). A concession algorithm determines the degree of concession for each negotiation round, and a tradeoff algorithm is required to generate a proposal in multi-issue negotiation. The tradeoff algorithm generates a proposal by combining proposals for individual issues. If the negotiation issues are price and response time, a proposal is a combination of price and response time (e.g., low price with slow response, or high price with fast response). Unlike existing mechanisms can make only one proposal at a time, in this study, agents are allowed to make multiple proposals concurrently in a round that generated the same aggregated utility (i.e., 'burst proposal' [16]), differing only in terms of individual utilities.

The utility function $U(x)$ represents an agent's level of satisfaction with negotiation outcome x (e.g., $U(P)$ for price). For a decision-making, agents evaluate proposals according to the utility function. To define a price utility function, the negotiator needs to specify the most and the least preferred price. In general, the range of the utility function is $\{0\} \cup [u_{\min}, 1]$, where $U(P) = u_{\min}$ and $U(P) = 1$ represents the least and the most preferred price, respectively.

The time slot utility function defined in [16] supports participants in representing the temporal preferences for leasing/lending services. A consumer can specify the time slot utility function according to his/her work schedule, and a CSP can specify the time slot utility according to the expected resource demands at any given time. CSPs may charge a higher price at peak time and a lower price at off peak, and consumers may need to pay a higher price to use a service in more desirable time slots. Fig. 3 shows an example of generated time slot utility function. The consumer who uses this function will have the highest utility at 15T.

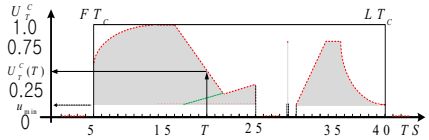

Figure 3. Example of generated time slot utility function.

The service response time represents the minimum response time that a CSP offers. Let the initial response time (IRT) and reserve response time (RRT) be the most and least preferred response time, respectively. The response time given to a consumer can be evaluated by the response time utility function of a consumer, as follows:

$$U_{RT}^{C}(RT) = \begin{cases} u_{min} + (1 - u_{min}) \cdot \left| \dfrac{RRT_C - RT}{RRT_C - IRT_C} \right|, & IRT_C \le RT \le RRT_C \\ 0, & \text{o.w.} \end{cases} \quad (1)$$

Finally, the aggregated utility function, which includes service price, time slot, and response time, is as follows:

$$U_{Total}(P, TS, RT) = \begin{cases} 0, & (U_P = 0,\ U_{TS} = 0, or\ U_{RT} = 0) \\ w_P \cdot U_P + w_{TS} \cdot U_{TS} + w_{RT} \cdot U_{RT}, & \text{o.w.} \end{cases} \quad (2)$$

Fig. 4 shows empirical results of the proposed SLA negotiation in terms of negotiation speed and utility using an agent-based Cloud testbed [16]. The proposed burst mode (B10, B50, and B100) and the adaptive burst mode (AB) achieved a higher average total utility and a faster agreement speed than related schemes (middle: M1, random: R1, and heuristic: H1[12]) that can generate only one proposal in each negotiation round.

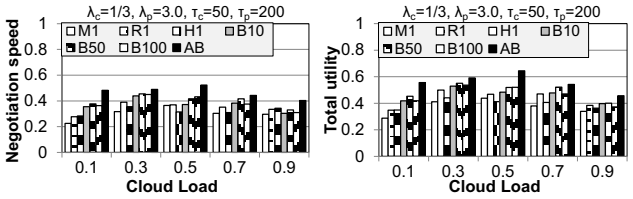

Figure 4. Simulation results: effect of the proposed trade-off algorithm.

C. SLA-driven resource allocation in gloval data centers

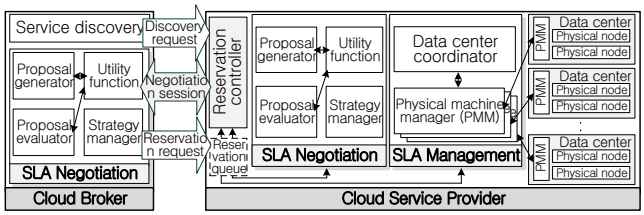

Figure 5. SLA negotiation and management based Cloud framework.

The proposed framework [17] consists of a service broker and CSP (Fig. 5). The broker connects a consumer to the CSP who owns the service discovered and has SLA-E capability through the SLA negotiation component. A CSP

consists of (1) reservation controller, (2) SLA negotiation, (3) SLA-M, and (4) distributed data centers. The SLA-M component, which is called the workload- and location-aware resource allocation (WLARA), selects a center among the global data centers to allocate the requested service. The conditions (i.e., utility-based evaluation [17]) of selecting a data center are based on workload and the SLA (service response time in this work). Each data center includes a physical machine manager, who manages the physical computing nodes of a data center to evaluate the average response time of a data center. Using the monitoring, SLA-M selects a data center and specific physical computing node.

Fig. 6 shows the performance of WLARA and other schemes in terms of SLA violations and placement failures [17]. Fig. 6(a) shows agreed and measured response time in WLARA. Consumers have different response time thresholds according to the outcomes of the negotiated SLA. In Fig. 6(b), with WLARA, the least number of SLA violations is guaranteed, whereas the greedy, random, RR, NIM, and IM (a similar way with EC2) schemes caused more violations because WLARA considers both workload and response time (including network delay) in a utility function. Hence, WLARA can allocate a consumer's request to a data center that has a lower workload and is physically closer to guarantee the response time threshold in the SLA.

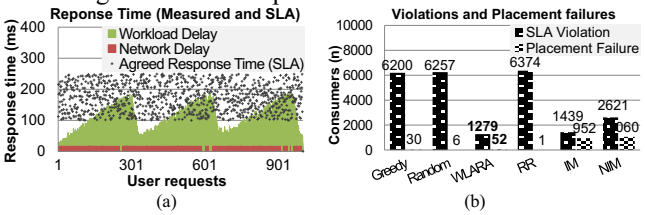

Figure 6. Simulation results: SLA-driven resource allocation.

V. Conclusions and Future Work

The novelty and significance of this study are 1) the design of a multi-issue negotiation mechanism that facilitates the price, time slot, and QoS negotiation for SLA-E, and 2) the development of a SLA-driven Cloud framework that includes the SLA negotiation mechanism and a workload- and location-aware resource allocation to manage SLAs.

The expected contributions of this research are as follows: 1) while the variety of SLA options is limited within enforced SLA strategies, the different preferences of a consumer and CSP can be narrowed efficiently through the proposed SLA negotiation; 2) the time slot negotiation can provide a market-based pricing scheme, and we observed that the proposed mechanism as a pricing scheme has advantages over the pricing schemes used in EC2 [16]; 3) the design of tradeoff algorithms considers the tradeoff relationship among utilities to enhance utility and negotiation speed. Also, to prevent SLA violations, 4) we provide a SLA-driven resource allocation that selects a data center among globally distributed data centers operated by a CSP.

Finally, the authors expect this work can be extended in two ways: 1) considering and specifying additional negotiation issues in Cloud SLAs and 2) deploying the proposed system on a real infrastructure and evaluating the performance with real workloads ([16] includes a case study).

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST; No. 2010-0026438) and by PLSI supercomputing resources of the Korea Institute of Science and Technology Information. Thanks to Prof. Kwang Mong Sim for his valuable advice.

References

[1] A. Andrieux, K. Czajkowski, A. Dan, K. Keahey, H. Ludwig, T. Nakata, J. Pruyne, J. Rofrano, S. Tuecke, and M. Xu, Web Services Agreement Specification (WS-Agreement), Open Grid Forum, 2006.

[2] R. Buyya, S. Garg, and R. Calheiros, "SLA-Oriented resource provisioning for cloud computing: challenges, architecture, and solutions," In *Proc. International Conference on Cloud and Service Computing*, pp. 1–10, 2011.

[3] K. M. Sim, "Grid resource negotiation: survey and new directions," *IEEE Trans. Syst., Man, Cybern. C, Applications and Reviews*, vol. 40, no. 3, pp. 245–257, May 2010.

[4] R. Lawley, M. Luck, K. Decker, T. R. Payne, and L. Moreau, "Automated negotiation between publishers and consumers of grid notifications," *Parallel Proc. Lett.,* vol. 13, no. 4, pp. 537–548, 2003.

[5] K. M. Sim, "G-commerce, market-driven G-negotiation agents and Grid resource management," *IEEE Trans. on Syst, Man and Cybern. B, Cybernetics*, vol. 36, no. 6, pp. 1381–1394, Dec. 2006.

[6] F. Lang, "Developing dynamic strategies for multi-issue automated contracting in the agent based commercial grid," in *Proc. IEEE Int. Symp. Cluster Comput. Grid (CCGrid05), Workshop Agent-Based Grid Econ.* Cardiff, U.K. pp. 342–349, May 2005.

[7] H. Gimpel, H. Ludwig, A. Dan, and R. Kearney, "PANDA": Specifying policies for automated negotiations of service contracts," in *Proc. ICSOC*, vol. 2910, LNCS, New York, pp. 287–302, 2003.

[8] S. Venugopal, X. Chu, R. Buyya, "A negotiation mechanism for advance resource reservation using the alternate offers protocol," in *Proc. 16th Int. Workshop on Quality of Service (IWQoS 2008)*, Twente, The Netherlands. June 2008.

[9] A. Rubinstein, "Perfect equilibrium in a bargaining model," *Econometrica*, vol. 50, no. 1, pp. 97–109, 1982.

[10] I. Brandic, D. Music, and S. Dustdar, "Service mediation and negotiation bootstrapping as first achievements towards self-adaptable grid and cloud services," In *Proc. Grids meet Autonomic Computing Workshop (GMAC) 2009*, Spain, June 2009.

[11] A. Paschke, J. Dietrich, K. Kuhla, "A Logic Based SLA Management Framework," in *Proc. Semantic Web and Policy Workshop (SWPW), 4th Semantic Web Conference (ISWC 2005)*, Galway, Ireland, 2005.

[12] J. Yan, R. Kowalczyk, J. Lin, M. B. Chhetri, S. K. Goh, and J. Zhang, "Autonomous service level agreement negotiation for service composition provision," *Future Generation Computer Systems*, vol. 23, no. 6, pp. 748–759, July 2007.

[13] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," *IEEE Internet Computing*, vol. 13, no. 5, pp. 14–22, 2009.

[14] R. Buyya, A. Beloglazov, and J. Abawajy, "Energy-efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges," In *Proc. the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA2010)*, pp. 6–20

[15] K. Le, J. Zhang, J. Meng, R. Bianchini, T. D. Nguyen, and Y. Jaluria, "Reducing electricity cost through virtual machine placement in high performance computing clouds," In *Proc. 2011 Super Computing (SC11)*, Washington, USA.

[16] S. Son and K. M. Sim, "A price and time slot negotiation mechanism for cloud service reservations," *IEEE Trans. on Systems, Man and Cybernetics, B: Cybernetics*, vol. 42, no. 3, pp. 713–728, June 2012.

[17] S. Son, G. Jung, and S. C. Jun, "An SLA-based cloud computing that facilitates resource allocation in the distributed data centers of a cloud provider," *The Journal of Supercomputing,* Jan. 2013. doi: 10.1007/s11227-012-0861-z.