Rule-based SLA Management for Revenue Maximisation in Cloud Computing Markets

Mario Macías, J. Oriol Fitó and Jordi Guitart

Barcelona Supercomputing Center and Universitat Politecnica de Catalunya
Jordi Girona 29
08034 Barcelona, Spain
{mario.macias, josep.oriol, jordi.guitart}@bsc.es

***Abstract*—This paper introduces several Business Rules for maximising the revenue of Providers in Cloud Computing Markets. These rules apply in both negotiation and execution time, and enforce the achievement of Business-Level Objectives by establishing a bidirectional data flow between market and resource layers. The experiments demonstrate that the revenue is maximized by using both resource data when negotiating, and economic information when managing the resources.**

I. Introduction

Cloud Computing [1] arisen as a successful computing paradigm, because it allows hiring resources without caring of maintenance costs and adds some new features for clients, such as the possibility of scale-up and scale-down resources dynamically depending on punctual requirements.

Within a Cloud ecosystem, a market-based approach would motivate the Providers to offer their resources in the system and give a Quality of Service (QoS) according to their real capacity. In addition, market mechanisms obligate the users to adjust their reservations to their real requirements. In a Cloud computing market, brokers that represent Service Providers and Clients negotiate for establishing the QoS terms within a Service-Level Agreement (SLA). The provider performs the negotiation and the enforcement of SLAs by pursuing its own Business-Level Objectives (BLO).

This paper encompasses the autonomic enforcement of a single BLO: the maximisation of the revenue. This paper defends the idea that revenue can be maximized by establishing a bidirectional data flow between market and resource layers: market brokers can perform negotiations that are more profitable if they use resource-level data, and the resource manager can help maximising the revenue if it manages the SLAs by considering this BLO. This bidirectional data flow is performed by an intermediate entity, called Economically Enhanced Resource Manager (EERM) [2].

Related work set out the necessity of managing the resources by considering the BLOs [3], [4]. However, most models are restricted to prioritize users who spend more in online shops. This paper deals with a heterogeneous scenario where workloads can be both Web Services or Batch Jobs.

The intention of this work is to provide an integrated set of policies that work together to maximise the profit of a Cloud provider, dealing also with performance issues. The introduced policies are evaluated in terms of relative results and tendencies, not in terms of absolute values.

The rest of the paper is structured as follows: after the definition of the scenario in section II. Section III describes the proposed rules for maximising the Revenue. Section IV describes the experiments and their results. At the end, the conclusions are summarized and future work is explained.

II. Previous definitions

Each EERM controls a set of physical machines. Each physical machine can host several Virtual Machines that can execute single tasks, such as Web Services or Batch Jobs. The QoS terms of a task are described in $SLA = \{Rev(vt), \overrightarrow{S}, \Delta t\}$, where:

- $\overrightarrow{S}$ describes the QoS of the purchased service.
- Δt is the time period requested for allocating the task.
- $Rev(vt)$ is the revenue function that describes the money that the provider earns after finishing correctly or incorrectly a task. vt is the amount of time in which the provider has not provided the agreed QoS to the client. Let MP be the maximum penalty and MR the maximum revenue, equation 1 describes the revenue function. If $vt < MRT$ the SLA is not violated (0 violations); if $vt > MPT$, the SLA is completely violated (1 violations). If $MRT > vt > MPT$, there is a partial violation ($\frac{vt-MRT}{MPT-MRT}$ violations).

$$Rev(vt) = \frac{MP - MR}{MPT - MRT}(vt - MRT) + MR \quad (1)$$

III. Description of the rules

A. Dynamic Pricing

In a market competition, a provider must establish rules for establishing variable prices in function of the offer/demand proportion. Previous work of the authors in this field established a formula that defined an aggressiveness factor as a function of the resources load status: the prices will be higher when the system workload is higher [5].

$$u_{rv}(\overrightarrow{S}) = 0.5 + \frac{\sin\left(\frac{\pi}{2}\left(2u_p(\overrightarrow{S}) + \left(1 - a(\Delta t)^{15}\right)\right)\right)}{2} \quad (2)$$

Each of the components of the equation 2 is thoroughly explained in the previous work [5]. $a(\Delta t)$ is calculated differently in this paper: instead of using current system data, the EERM uses future predictions about the resources load. Let $C_{used}(t)$ be a function that predicts the usage of the currently reserved resources over time in terms of CPU; let $C_{req}(t)$ be a constant function that represents the CPUs requested by the client in the negotiation process; let $C_j(t)$ be a constant function that represents the number of CPUs of the physical resource j; Equation 3 shows how the aggressiveness factor $a(\Delta t)$ is calculated from a set of N physical machines. It assumes that CPU is the bottleneck of the system, but it could be changed by other type of resource.

$$a(\Delta t) = \frac{\sum_{j=1}^{N} \int_{t_i}^{t_f} C_{used}(t) + C_{req}(t)\, dt}{\sum_{j=1}^{N} \int_{t_i}^{t_f} C_i(t)\, dt} \quad (3)$$

B. Resource Overprovisioning

If there are not enough unreserved resources at a given time, a classical RM will refuse a SLA proposal from the client. However, clients do not always use the total of resources that they have reserved, and these unused resources could be resold to other clients for increase the revenue.

Based on a prediction of the usage of resources at a given time slot ($C_{used}(t)$), the EERM uses the scores all the set $j = \{1 \dots N\}$ of physical resources as defined in equation 4. Finally, the physical resource j of which score is the higher positive is selected for executing the incoming task. If there are not resources with positive score, the job is rejected.

$$score_j = 1 - \frac{\int_{t_i}^{t_f} C_{used}(t) + C_{req}(t)\, dt}{\int_{t_i}^{t_f} C_i(t)\, dt} \quad (4)$$

C. Selective SLA Violation

When the provider is not able to fulfil all the SLAs that has agreed, the EERM can perform a selective violation of some SLAs for minimising the economic impact of the penalties [2]. The set of SLAs to violate is chosen according to the next process: the future profit of the provider is estimated for each possible SLA violation in the system, by adding all the revenues and penalties of all the SLAs. After all the possibilities are calculated, the resources of the SLA of which violation produces the higher gain (or the lower loss) are deallocated temporarily to leave free space for the other SLAs.

D. Selective SLA Cancellation

When the client starts a negotiation for a task that can not fit in the system due to space limitations, it is possible to cancel the tasks that are already scheduled or running in the system if the revenue of the incoming task is high enough to compensate for the penalty of the cancelled SLA.

This policy must be executed with caution, because the short-term benefit is in conflict with mid-term losses in the reputation of the provider [6]. Because this paper does not consider reputation, a cancellation policy is applied without restrictions: for each SLA of which time slot collides with the demand, and of which possible cancellation would free enough space to allocate the incoming SLA, the benefit of cancelling it is estimated by subtracting the maximum penalty of the violation to the maximum price that client could pay for the incoming task. The SLA of which cancellation reports the highest profit is marked as *cancellable* and the provider negotiates the maximum price with the client. If the provider accepts, the SLA is cancelled and the new task is allocated.

E. Ranges for Quality of Service

Different ranges of QoS entails different values for MRT, MR, MPT and MP in $Rev(vt)$. Three ranges of QoS have been defined, from higher to lower: Gold, Silver and Bronze. Higher QoS ranges have higher values for MR and lower for MP, and MRT and MPT are 0. The QoS range that establishes $Rev(vt)$ must be negotiated by client and provider. The combination of different QoS ranges and rules for selective SLA violation and cancellation lead to less violations of high-range QoS.

F. Tasks Reallocation

The heterogeneous nature of Cloud Computing tasks can lead to the unbalancement of workloads in the resources pool, thus some resources can become overloaded. For avoiding this problem, the EERM migrates VMs from overloaded physical resources to the less loaded resources.

Recent studies [7] reveal that the cost of migrating web services in Cloud Computing is near zero thanks to virtualization, because creating, booting, and populating a virtual machine with data requires few seconds (negligible in tasks of which duration is from one to several hours).

G. Redistribution of Assigned Resources

When calculating the allocation of plain resources in function to high-level QoS parameters, the SLA Decomposition process [8] have an associated error rate that can derive to future violations of SLAs, if the provider allocates insufficient resources, or to a waste of resources if it allocates more resources than needed.

Redistribution of resources can compensate the inaccuracy of SLA decomposition: the EERM will look for the tasks that are underutilizing their resources and, if there are enough underutilized resources, a sufficient portion of them will be unassigned from their current tasks, and assigned to the task with insufficient resources. This process is easy to implement thanks to Virtualization technology [9].

IV. Evaluation

A. Experimental Environment

The simulations consider two types of workload: Web Services and Batch Jobs. Web Services have time slots and

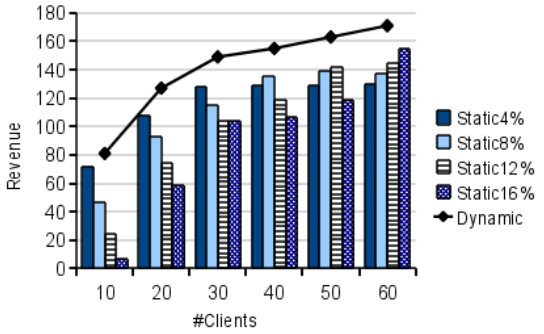

Figure 1. Comparison of revenue between dynamic and fixed pricing

the workload can vary over time. Batch Jobs have a stable CPU workload, and the time slot is variable because they do not have strict deadline requirements (they could be executed at early morning). The workload for Bath Job has a pseudo-random distribution and the workload for Web services have a variable workload distribution taken from a real Web application during one week.

To start a negotiation, the client sends a SLA proposal to the provider that specifies the time slot (fixed or variable), the required amount of resources and the QoS range. If the provider accepts the proposal, it returns the SLA by specifying the price. The client chooses the provider that, at equal QoS, offers the lower price.

For the same task in equal time and load conditions, Gold-QoS tasks have a Reservation Price for the seller 50% higher than the Bronze Reservation Price, and Silver tasks have a Reservation Price 20% higher than Bronze tasks.

B. Dynamic pricing

Five providers with 12 CPUs are competing in a market of which clients send Bronze, Silver and Gold tasks at the same proportion. There is one provider that implements dynamic pricing, and four providers that offer fixed prices, which are always a fixed proportion between the Reservation Price of the Seller and the Reservation Price of the Buyer, labelled from *Static4%* to *Static16%* in function of this proportion. The experiments are repeated with different demand levels.

Figure 1 shows the revenues of all the providers in the market. Comparing fixed-pricing providers with the dynamic-pricing provider, it can be seen how dynamic-pricing always gets the highest revenue in all the demand levels, because it can adapt better to all the possible scenarios.

C. Resources overprovisioning

In this experiment, two providers are competing in a Market. The first provider performs Dynamic Pricing and Overprovisioning, and the second only performs Dynamic Pricing. Both providers manage two 8-CPU servers that will host the Virtual Machines where the tasks will be executed. The experiments are repeated with several demand levels. The size of the virtual machines can vary from 1 to 4 CPUs, in function of the hour of day (1 CPU in off-peak hours, 4 CPU in peak hours).

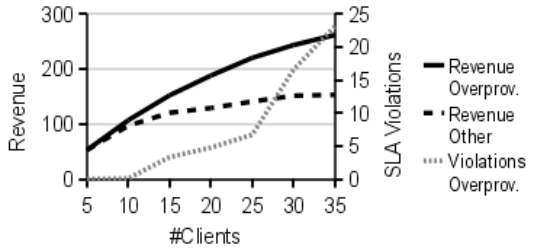

Figure 2. Comparison of revenue and SLA violations with and without Resource Overprovisioning

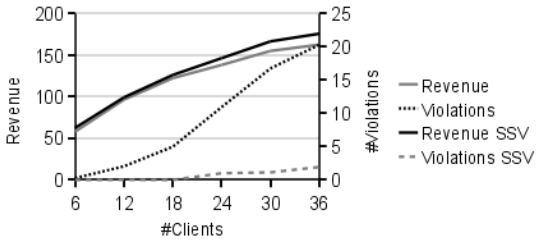

Figure 3. Providers with selective SLA violation maximize their benefit

The predictor component that calculates $C_{used}(t)$ has an error rate of 10%. According to current research, it is a reasonable rate [8].

Figure 2 shows how the revenue of a provider does not grow linearly in function of the demand, because the resources are finite. However, a provider with overprovisioning can allocate more tasks and its benefit is higher. The drawback of resource overprovisioning is that some SLAs are violated, although the number of violations represent the $\sim 1\%$ of the total.

Extended experiments in related work demonstrate that the scoring function of Equation 4 indirectly allocates tasks with variable time requirements in slots with low demand [10].

D. Combining Ranges for Quality of Service and Selective SLA Violation

Two providers with 8 CPUs are competing in a market. Both perform Dynamic Pricing and Overprovisioning, but only one implements Selective SLA Violation. The experiments are repeated with a variable number of clients that demand different ranges of QoS (Gold, Silver, and Bronze).

Figure 3 shows how the provider that implements Selective SLA Violation earns between 5-10% more money than the other provider. The number of violations is $\sim$90% lesser with Selective SLA Violation, because the EERM focuses the violations in those SLAs of which $vt < MRT$.

The difference between the QoS ranges is how MRT, MR, MPT and MP are located. The experiments of this paper use the next values: $MRT(Bronze, Silver, Gold) = (15\%, 5\%, 3\%)$, $MPT(Bronze, Silver, Gold) = (75\%, 50\%, 30\%)$ and $MP(Bronze, Silver, Gold) = (MR, 2MR, 3MR)$. These values are arbitrary, but they allow to show that Gold clients are less allowed to have violations than Silver, and Silver less than Bronze, high QoS ranges are economically less

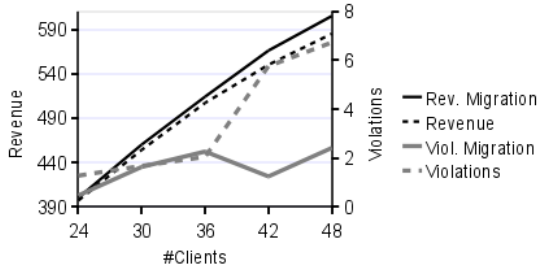

Figure 4. Providers that implement service migrations minimise SLA violations and increase their revenue

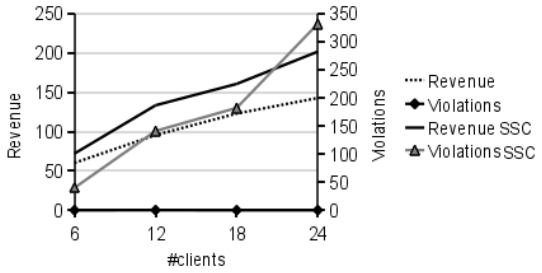

Figure 5. Selective SLA cancellation increases greatly the revenue, but also the violations of SLAs

permissive with violations. Changing these values would not alter qualitatively the simulation results, which shown that Bronze SLAs have a failure rate up to 400% higher than Gold SLAs, and 200% higher than Silver SLAs.

E. Tasks Reallocation

Two providers with 4 physical machines of 8 CPUs each one are competing in a market. Both have the policies already tested in this section. Only one provider performs tasks reallocations. Figure 4 shows how dynamic reallocation on machines of which workload is $>$80% increases the revenue up to 12% and minimises the number of violations in high-demand scenarios.

F. Selective SLA cancellation

Two providers with a single machine of 8 CPUs are competing in a market. Both providers are configured to perform Dynamic Pricing, Resources Overprovisioning and Selective SLA Violations. Only one performs Selective SLA Cancellations. The same experiment is repeated with different number of clients.

Figure 5 shows that the revenue is increased a $\sim$50% by applying Selective SLA cancellation, but also the violation of SLAs. If market provided a reputation system, this policy would not be valid. However, it can be applied in certain situations, such as arrival of tasks from a special user, re-organisation after partial failures of the system, etc.

G. Redistribution of allocated resources

In this scenario, two providers with 16 CPU are competing in a services market. Both providers implement dynamic pricing and resource overprovisioning, but only one implements dynamic resource redistribution. The results demonstrate that the provider that performs resource redistribution violates less SLAs. However, if the market does not provide a reputation system there is not an important difference in revenue.

V. Conclusions and Future Work

This paper proposes the improvement of SLA Negotiation and Management in Cloud Computing markets by means of bidirectional communication between market brokers and resource managers. It also introduces several Rules for maximising revenue in a Cloud Provider, and demonstrates their validity by means of several experiments.

This paper be continued through two future research lines: the creation policies for achieving other BLOs, such as client classification, and the addition of support for dynamic rules that can be automatically modified by the EERM for allowing a better adaptation to changing market environments.

Acknowledgements

This work is supported by the Ministry of Science and Technology of Spain and the European Union (FEDER funds) under contract TIN2007-60625, by the Generalitat de Catalunya under contract 2009-SGR-980, and by the European Commission under FP7-ICT-2009-5 contract 257115 (OPTIMIS).

References

[1] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities," *High Performance Computing and Communications, 10th IEEE International Conference on*, vol. 0, pp. 5–13, 2008.

[2] M. Macías, O. Rana, G. Smith, J. Guitart, and J. Torres, "Maximizing revenue in Grid markets using an Economically Enhanced Resource Manager," *Concurrency and Computation: Practice and Experience*, p. n/a, September 2008. [Online]. Available: http://dx.doi.org/10.1002/cpe.1370

[3] D. A. Menascé, V. A. F. Almeida, R. Fonseca, and M. A. Mendes, "Business-oriented resource management policies for e-commerce servers," *Perform. Eval.*, vol. 42, no. 2-3, pp. 223–239, 2000.

[4] N. Poggi, T. Moreno, J. Berral, R. Gavalda, and J. Torres, "Self-adaptive utility-based web session management," *Computing Networks*, vol. 53:10, pp. 1712–1721, 2009.

[5] M. Macías and J. Guitart, "Using resource-level information into nonadditive negotiation models for cloud market environments," in *12th IEEE/IFIP Network Operations and Management Symposium (NOMS'10)*, Osaka, Japan, April 2010, pp. 325–332.

[6] M. Macías and J. Guitart, "Influence of reputation in revenue of grid service providers," in *2nd International Workshop on High Performance Grid Middleware (HiPerGRID 2008)*, Bucharest, Romania, November 2008.

[7] I. Goiri, F. Julia, and J. Guitart, "Efficient data management support for virtualized service providers," in *17th Euromicro Conference on Parallel, Distributed and Network-based Processing (PDP'09)*, Weimar, Germany, February 2009, pp. 409–413.

[8] G. Reig, J. Alonso, and J. Guitart, "Prediction of job resource requirements for deadline schedulers to manage high-level slas on the cloud," in *9th IEEE International Symposium on Network Computing and Applications (NCA'10)*.

[9] I. Goiri, F. Julia, J. Ejarque, M. de Palol, R. Badia, J. Guitart, and J. Torres, "Introducing virtual execution environments for application lifecycle management and sla-driven resource distribution within service providers," in *8th IEEE Intl. Symposium on Network Computing and Applications (NCA'09)*, Massachusetts, USA, July 2009, pp. 211–218.

[10] M. Macías and J. Guitart, "Maximising revenue in cloud computing markets by means of economically enhanced SLA management," Computer Architecture Department, Universitat Politecnica de Catalunya, Tech. Rep. UPC-DAC-RR-CAP-2010-22, September 2010. [Online]. Available: http://gsi.ac.upc.edu/reports/2010/32/policydescription.pdf