A Game-theoretic Framework for Creating Optimal SLA/Contract

Sofia Moroni, Nicolas Figueroa, Alejandro Jofre, Akhil Sahai, Yuan Chen, Subu Iyer
Enterprise Systems and Software Laboratory
HP Laboratories Palo Alto

HPL-2007-126
July 26, 2007*

game theory, SLA, framework, contract

An SLA/Contract is an agreement between a client and a service provider. It specifies desired levels of service and penalties in case of default. It is of interest from the Service Providers point of view, to determine the optimal contract, that will maximize its utility. In this work we model the situation based on the notion of Moral Hazard: providing a good service is costly and results are affected by the resources involved. As a consequence, a credible contract must fulfill the incentive compatibility constraint. We extend the above model to take into account the possibility that there might different types of clients, and that the Service Provider will offer a menu of contracts intended for each of these clients, as a means of maximizing utility. From the Service Providers point of view, finding an optimal contract will consist of solving a nonlinear optimization problem subject to constraints. We derive conditions under which these constraints will take a simple form and we analyze a scenario, in which, the randomness comes from the Response Time of a given IT Service, and the input is the number of servers that will be dedicated to each client.

* Internal Accession Date Only

Approved for External Publication

A Game-theoretic Framework for Creating Optimal SLA/Contract

Sofia Moroni[a,0], Nicolas Figueroa[a,0], Alejandro Jofre[a,0], Akhil Sahai[b], Yuan Chen[b], and Subu Iyer[b]

[a] *Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Beauchef 850, Santiago, Chile.*
[b] *HP Laboratories, Palo-Alto, CA 94034.*

Abstract

An SLA/Contract is an agreement between a client and a service provider. It specifies desired levels of service and penalties in case of default. It is of interest from the Service Providers point of view, to determine the optimal contract, that will maximize its utility. In this work we model the situation based on the notion of Moral Hazard: providing a good service is costly and results are affected by the resources involved. As a consequence, a credible contract must fulfill the incentive compatibility constraint. We extend the above model to take into account the possibility that there might different types of clients, and that the Service Provider will offer a menu of contracts intended for each of these clients, as a means of maximizing utility. From the Service Providers point of view, finding an optimal contract will consist of solving a nonlinear optimization problem subject to constraints. We derive conditions under which these constraints will take a simple form and we analyze a scenario, in which, the randomness comes from the Response Time of a given IT Service, and the input is the number of servers that will be dedicated to each client.

[0] Work developed during stay at HP Labs Palo-Alto

Contents

1. Introduction

An SLA/Contract is an agreement between a provider and a consumer, it is comprised of Service Level Objectives that guarantee quality of service, such as availability, performance and reliability, a promise of payment and penalties to impose in case of violation and other characteristics (such as validity period, scope and restrictions).

This work will focus on how to find the optimal contract from the Service Provider's (SP) point of view. We model SLA/Contracts using the concept of Moral Hazard and Adverse Selection. The intent is to maximize the SP's utility while keeping the customer's utility in mind.

The basic idea is that the provider, through some costly effort (investment, use of scarce resources such as number of CPU's, number of engineer hours, etc.), can increase the quality of the service, but that there is also an additional stochastic component to it. In this hidden action we identify the concept of Moral Hazard. Greater levels effort increase the probability of a good quality outcome. In the context of the services that an IT provider, better infrastructure on average provides better performance, but some unforeseen incidents(extra demand, breakdown of a system, etc.) may still lead to poor quality. The client can observe a quality, but not the effort. Therefore, the only way to induce a high level of effort is through a compensation system that is "steep"m i.e. payments will be higher when observed quality is better. Nonetheless, this affects the provider, since she may sometimes be punished for low quality even if the effort put in the process was high. Of course, sometimes she will be rewarded in excess for her effort. The basic trade-off is then set: "steeper" compensation systems will induce higher effort, but they will shift more risk (in terms of earnings) to the provider, who is risk averse and will charge more for the service.

2. Theoretical Model

2.1 Basic Model

The provider delivers a service quality $q_m \in Q_M \subseteq I\!R$ (measured in monetary terms) to the client. This quality level depends stochastically on the effort level $e \in E \subseteq I\!R^{n_e}$ that the provider invests in the client. Clients will not be able to observe the level of effort that the SP has assigned to each one of them. The service quality in monetary terms q_m will depend on an observable and verifiable variable $q \in Q \subseteq I\!R^{n_q}$ that we will call the (true) quality of service. We suppose that there is a function $g : Q \rightarrow Q_M$ that takes the variable q and assigns to it the monetary value of the quality of service q_m, which will represent how the client values quality. g will be non-decreasing and concave in each one of the components of q. We assume that the distribution function of the (true) quality, given a level of effort e is $f_q(q|e)$ and with cumulative function $F_q(q|e)$, and its support will be a cube in $I\!R^{n_q}$, $\bigotimes_{p=1}^{n_q}[\underline{q_p}, \bar{q_p}]$, independent of e.

Since the level of effort is not observable by the client, a contract can only specify a payment contingent on the quality, that is observable and verifiable. We assume that the payment will be a function of monetary quality of service, which will be observable since it is a deterministic function of quality. We will denote this payment rule as $\bar{p}(q) = p(g(q)) = p(q_m)$. It is important that the q_m variable is single dimensional, because otherwise the problem becomes much more complex.

The client has a utility function V, so for a realized level of quality, his utility will be $V(q_m - p(q_m))$. The provider, on the other hand, has a utility function that depends on effort and money, so for a given level of effort e and quality q_m, her utility is $U(p(q_m), e)$. We assume $V' > 0$ and $V'' \leq 0$ and that $\frac{\partial U}{\partial p} > 0$, $\frac{\partial U}{\partial e} < 0$, $\frac{\partial^2 U}{\partial^2 p} \leq 0$ and $\frac{\partial^2 U}{\partial^2 e} \geq 0$ (see figure 1). Then the provider chooses e and $p(q_m)$ to solve

$$\max_{p(),e} \int U(p(q_m), e) f_{q_m}(q_m|e) dq_m \tag{1}$$

subject to

$$\int V(q_m - p(q_m)) f_{q_m}(q_m|e) \geq \bar{V} \tag{2}$$

$$e \in \text{argmax}_{e'} \int U(p(q_m), e') f_{q_m}(q_m|e') dq_m \tag{3}$$

The objective is to maximize the expected utility of the service provider. (2) (called participation constraint, from now on "PC"), states that the utility level of the client has to be above a certain level $\bar{V}$. This level reflects the opportunity cost of the resources involved, and for this simple model will be considered a parameter. The constraint (3) is now a *credibility constraint*. The promised effort level has to be optimal given the contract, since otherwise the client would not trust a contract with such a promised effort level.

In terms of q this optimization can be written equivalently as:

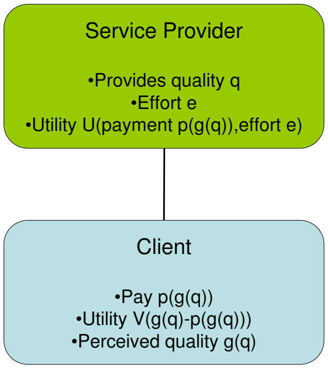
$$\max_{p(),e} \int U(p(g(q)), e) f_q(q|e) dq \tag{4}$$

subject to

$$\int V(g(q) - p(g(q)))f_q(q|e)dq \geq \bar{V} \tag{5}$$

$$e \in \text{argmax}_{e'} \int U(p(g(q)), e')f_q(q|e')dq \tag{6}$$

Figure 1: One Client Model



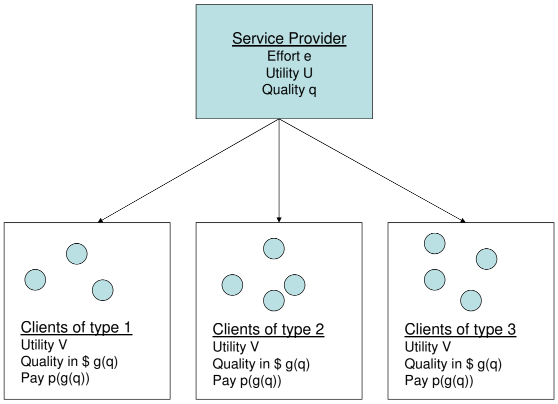
2.2 The General Model

We can enrich the model to take into account that the SP could have different clients, that can be classified in "types". The agents that belong to a particular type will differ from the other types because they can have different valuations of the service, and utility functions. The Service Provider will offer a menu of contracts, to satisfy the different necessities of the clients and therefore extract more payments.

The main constraint is that Service Provider does not know which client is which, since the particular valuation for the service is private information of the client, or that he has to offer all contracts to all clients for legal reasons. Given this constraint, the contracts have to satisfy a *self-selection* property: each client selects the contract that was designed for his particular type because is the one that gives him the greatest utility level.

For the model, let's suppose that there are N clients, which can be classified in k types. We denote by θ_j the "type" of agents in the class j, and we assume that there is an amount μ_j of those agents. The Service Provider will offer a menu of contracts, each of them designed for one type of client. As before the contracts will offer a payment given the realized monetary quality q_m^i that client i receives. We will denote the contract intended for agents of type θ_j as $p_{\theta_j}(q_m)$. The Service Provider will devote independent efforts for each client, which we will denote $\{e_{i,\theta_j}\}_{i\in\{1,\dots\mu_j\},j\in\{1\dots k\}}$, that is, the effort that the i'th client of type θ_j will be assigned, will be e_{i,θ_j}. The utility function of the Service Provider will take the form $U(\{p_{\theta_j}(\cdot)\}_j, \sum_i e_i)$, that is, it will depend on payments and the sum of efforts dedicated to each client, with $\frac{\partial U}{\partial p_i} > 0$, $\frac{\partial U}{\partial e_i} < 0$, $\frac{\partial^2 U}{\partial^2 p_i} \leq 0$ and $\frac{\partial^2 U}{\partial^2 e_i} \leq 0$ (see figure 2)

Figure 2: Multi-Client Model



Clients of the same type, say θ_j will have the same utility function $V(\cdot|\theta_j)$, the same g_{θ_j} function and reservation utility $\bar{V}(\theta_j)$, that will relate true quality to monetary quality of service, and distribution function of quality $q_{\theta_j} \in \bigotimes_{p=1}^{n_{q_{\theta_j}}} [\underline{q}_{p,\theta_j}, \bar{q}_{p,\theta_j}]$ contingent on effort, $f_{\theta_j}(q_{\theta_j}|e)$. The distribution function of monetary quality will depend on the effort level that each client receives, and will be independent of the effort given to other clients. The optimization problem that the Service Provider will face is,

$$\max_{\{p_{\theta_j}(\cdot)\}_j, \{e_{i,\theta_j}\}_{j,i\in\{1,\ldots,\mu_j\}}} \int U(\{p_{\theta_j}(g_{\theta_j}(q^{i,\theta_j}))\}_{i,j}, \sum_{i,j} e_{i,\theta_j}) \prod_{j,i} f_{\theta_j}(q^{i,\theta_j}|e_{i,\theta_j}) d\vec{q} \tag{7}$$

subject to

$$\int V(g_{\theta_i}(q) - p_{\theta_i}(g_{\theta_i}(q))|\theta_i) f_{\theta_i}(q|e_{r,\theta_i}) dq \geq \int V(g_{\theta_i}(q) - p_{\theta_j}(g_{\theta_j}(q))|\theta_i) f_{\theta_i}(q|e_{t,\theta_j}) dq \qquad \forall j \neq i, \forall r, t \tag{8}$$

$$\int V(g_{\theta_i}(q) - p_{\theta_i}(g_{\theta_i}(q))|\theta_i) f(q|e_i) \geq \bar{V}(\theta_i) \tag{9}$$

$$\{e_{i,\theta_j}\}_{j,i\in\{1,\ldots,\mu_j\}} \in \text{argmax}_{\{e'_{i,\theta_j}\}} \int U(\{p_{\theta_j}(g_{\theta_j}(q^{i,\theta_j}))\}_{i,j}, \sum_{i,j} e'_{i,\theta_j}) \prod_{j,i} f_{\theta_j}(q^{i,\theta_j}|e'_{i,\theta_j}) \tag{10}$$

As before the objective is to maximize the expected utility of the service provider. (8) is the *Self Selection Constraint*, it states that in the optimal contract each client type will prefer their own contract to the ones intended for other client types. (9) and (10) are the *Participation Constraint* and *Credibility Constraint*, respectively.

The optimization problem presented above is complicated. If we introduce a number of assumptions on the utility functions and the distributions of quality given the efforts the framework will be much simplified (details in appendix).

For simplicity, in this work we focus in optimizing over functions $p(\cdot)$ that are linear. That is, we are looking for contracts that are linear in monetary quality of service. This contracts are simple and could easily be implemented in real applicatons. However all the results concerning FOA can also be applied if we are optimizing over a subset containing concave functions. With the assumptions made (7)-(10) and it will become,

$$\max_{\{p_{\theta_j}(\cdot)\}_j,\{e_{\theta_j}\}_j} \int U(\{p_{\theta_j}(g_{\theta_j}(q^{\theta_j}))\}_j, \sum_j \mu_j e_{\theta_j}) \prod_j f_{\theta_j}(q^{\theta_j}|e_{\theta_j}) \tag{11}$$

subject to

$$\int V(g_{\theta_i}(q) - p_{\theta_i}(g_{\theta_i}(q))|\theta_i) f_{\theta_i}(q|e_{\theta_i}) dq \geq \int V(g_{\theta_i}(q) - p_{\theta_j}(g_{\theta_j}(q))|\theta_i) f_{\theta_i}(q|e_{\theta_j}) dq \quad \forall j \neq i \tag{12}$$

$$\int V(g_{\theta_i}(q) - p_i(g_{\theta_i}(q))|\theta_i) f(q|e_i) \geq \bar{V}(\theta_i) \tag{13}$$

$$\{e_{\theta_j}\}_j \in \text{argmax}_{\{e'_{\theta_j}\}} \int U(\{p_{\theta_j}(g_{\theta_j}(q^{\theta_j}))\}_j, \sum_j \mu_j e'_{\theta_j}) \prod_j f_{\theta_j}(q^{\theta_j}|e'_{\theta_j}) \tag{14}$$

That is, all clients of the same type will have the same amount of effort assigned and the expected utility of the provider will be simplified to an integral over $\sum_{j=1}^{k} n_{q_{\theta_j}}$ variables.

3. Non-Bernoulli Utility Functions

In the framework outlined in previous sections we assumed that the utility of the SP and the client could be computed as an expected utility. Now we want to allow for the possibility that the utility of the agents doesn't have that form, that is, it is not a Bernoulli Utility Function. Our framework will remain the same, except that now the utilities of the agents will not be written as an expected utility, and they will be a function of the random variable of profits, that will, in turn, depend on the level of effort. We present two examples of non Bernoulli utilities, one of them is the Exp-Var and the other one is related to a the concept of the CVaR, used in finance to quantify risk. Both functions will depend on the expected profits and a measure of risk. Given a random variable X, that describes the behavior of uncertain profits, and with pdf $f_X(x)$ and cdf $F_X(x)$.[1]

Example 1 (Exp-Var Utility Function)

Definition 1 *The Exp-Var Utility function is defined as*

$$\tilde{U}(X) = \mathbb{E}(X) - \tau \int (x - \mathbb{E}(X))^2 f_X(x) dx.$$

The first term is the expected profits, τ is a constant and the second term is the variance of profits. The second term implies that the agents dislike risk, be it the SP or the clients, because it is a measure of how volatile profits can be. The Exp-Var function has the following advantageous property, in the context of the multi-dimensional effort,

Property 2 *If $g(q) = \sum_{i=1}^{n_q} g_i(q_i)$, $X = p(g(q)) - \phi(e)$, with $p(\cdot)$ linear, and $f(q|e) = \prod_{i=1}^{n_q} f_i(q_i|e)$ then*

$$\tilde{U}(p(g(\cdot)), e) = \sum_{i=1}^{n_q} \left(\mathbb{E}(p(g_i(q_i))) - \tau \int \left(p(g_i(q_i)) - \mathbb{E}(p(g_i(q_i)))\right)^2 f_i(q_i|e) dq_i \right) - \phi(e)$$

where $\mathbb{E}(p(g_i(q_i))) = \int p(g(q_i)) f_i(q_i|e) dq_i \quad \forall i$

That is, if we assume X is the profit from one client and the value of quality is a sum in each component of the quality, and those components have independent distributions, the utility will be "separable as a sum" in the sense that is outlined above. Note also that in the many client framework if X is the sum of the sum of payments from all clients, minus the total cost we will have a similar separability property. The advantage of this property is that, since each term of the sum consists on a one dimensional integral, sufficient conditions for FOA are more easily to be verified.

Example 2

Definition 3 (Value at risk) *For any level $\alpha \in (0,1)$ the value at risk, $VaR_\alpha(X)$ as*

$$VaR_\alpha(X) = -inf\{z | \mathbb{P}\{X \leq z\} > \alpha\}.$$

[1] In our framework $X = \bar{p}(q|e) - \phi(e)$, where ϕ is convex in e.

Definition 4 (Lower α-tail of X) *The lower α-tail of a r.v. X is the random variable X_α with distribution function*

$$F_{X_\alpha} = \frac{min\{\alpha, F_X\}}{\alpha}.$$

Definition 5 (CVaR, conditional value-at-risk) *For any $\alpha \in (0,1)$ the conditional value-at-risk, CVaR is defined as*

$$CVaR_\alpha(X) = -\mathbb{E}(X_\alpha).$$

where X_α is the lower α tail of X.

The utility function that we propose is the following,

$$\tilde{U}(X) = \mathbb{E}(X) - \tau CVaR_\alpha(X). \tag{15}$$

The CVaR is a measure of risk. For example, if $F_X(\cdot)$ is continuous, the CVaR will be the mean of the lowest $\alpha\%$ of the profits. If the SP has N clients, and we assume that the payments from each one of them is contingent on quality and therefore also a random variables that we will call $\{P_i\}_{i=1}^N$, then a possible utility function for the service provider would be,

$$\tilde{U}(\{P_i\}_{i=1}^N) = \sum_{i=1}^{N} (\mathbb{E}(P_i) - \tau CVaR_\alpha(P_i)) - \phi(e)$$

where $\phi(\cdot)$ is a convex function and represents the costs of effort and e summarizes the efforts assigned to all clients. In Property 15 we present conditions under which FOA will be valid for this particular type of non-Bernoulli utility function.

4. First Order Approach

Solving (4)-(6) is, in general, difficult, because of the last constraint. However, under certain conditions, if the maximization problem in (6) has an interior solution and the solution is unique, the last constraint can be simplified to an equality constraint which corresponds to the First Order Condition of the Optimization Problem in (3).[2] We will assume that efforts are chosen from an open set and that therefore we will only have to prove uniqueness of a stationary point[3]. This will happen, in particular, when the expected utility of the SP is concave in effort at the optimal contract.

This is known as the first order approach (FOA), and the conditions under which it is valid will depend on U and $f_q(q|e)$. In this section we study conditions that allow us, in the context of SLAs, to apply such a method.

A sufficient condition to be able to use the FOA is that the objective function be concave in e. If q is uni-dimensional we have the following results, From Jewitt (1988),

Property 6 *If $F_q(q|e)$ satisfies (16)-(17), then, for every function $\tilde{u}(\cdot) : I\!R \rightarrow I\!R$ in C^1 concave and non-decreasing, $\int \tilde{u}(q) f_q(q|e)dq$, will be concave in e and therefore, if $U(p(q_m), e) = u(p(q_m)) - \phi(e)$, with u concave and $\phi(\cdot)$ convex in e, then the optimization problem (1)-(3), solved for $p(\cdot)$ of linear form with positive slope, will satisfy FOA.*

$$\int_{-\infty}^{y} F_q(q,e)dq \textit{ is nonincreasing convex in } e \textit{ for each value of } y \tag{16}$$

$$\int_{-\infty}^{\infty} q f_q(q,e)dq \textit{ is nondecreasing and concave in } e \tag{17}$$

Proof. See Appendix ■

Corollary 7 *If $F_q(q|e)$ is convex in e for each q, then, if $U(p(q_m), e) = u(p(q_m)) - \phi(e)$ with u concave and $\phi(\cdot)$ convex in e, $\int u(p(g(q)))f_q(g(q)|e)dq$, will be concave in e, and therefore, the optimization problem (1)-(3), solved for $p(\cdot)$ of linear form with positive slope, will satisfy FOA.*

This results can be extended to the case in which we have multi-dimensional q, if some assumptions are made. If the Service Provider is risk neutral, the FOA approach can be used imposing conditions similar to (16)-(17) per component, without imposing any further conditions over the form of $f_q(q|e)$. Let $q = (q_1, q_2, \ldots, q_{n_q})$.

Property 8 *If the Service Provider is risk neutral, that is, $U(x, e) = x - \phi(e)$, and $g = \sum_{i=1}^{n_q} g_i(q_i)$, then if for every i, $F_{q_i}(q_i|e) = \int \int_{-\infty}^{q_i} f_q(q|e)dq_i dq_{-i}$ satisfies (16)-(17), then the optimization problem (1)-(3), solved for $p(\cdot)$ of linear form with positive slope, will satisfy FOA.*

[2] Note that this applies also for the case in which the set of efforts is not $I\!R^{n_e}$: the first order condition will correspond to the gradient of a Lagrangian.

[3] An stationary point is one in which the gradient of the SP's utility with respect to e becomes 0

If the SP is not risk neutral, further assumptions have to be made on the form of the distribution function. In general we have to require that $\int U(p(g(q)), e')f_q(q|e')dq$ be concave in e, or that it has a unique stationary point (since we are assuming an interior solution).

Property 9 *If* $U(p(g(q)) = \left(\prod_{i=1}^{n_q} \tilde{u}_i(q_i)\right)\phi(e)$*, with* $\tilde{u}_i(\cdot) > 0$ *concave and* $\phi(e) > 0$ *concave,* $f_q(q|e) = \prod_{i=1}^{n_q} f_{q_i}(q_i|e)$*, and (16)-(17) are satisfied for each* $F_{q_i}(q_i|e)$*, then the optimization problem (1)-(3), solved for* $p(\cdot)$ *of linear form with positive slope, will satisfy FOA, in terms of the* $\ln(U)$ *.*

If the utility function is separable as a multiplication but is now negative, we will need further assumptions to use FOA. However, we have to ask for a much stronger condition, which is that the logarithm of the utility be convex in e for every contract.

Property 10 *If* $U(p(g(q)), e) = -\left(\prod_{i=1}^{n_q} u_i(q_i)\right) \cdot \phi(e)$ *with* $u_i(\cdot) > 0$ *convex functions and* $\phi(e) > 0$ *convex,* $f_q(q|e) = \prod_{i=1}^{n_q} f_{q_i}(q_i|e)$*, then the optimization problem (1)-(3), solved for* $p(\cdot)$ *of linear form with positive slope, will satisfy FOA if* $\ln\left(\int u_i(q_i)f_{q_i}(q_i|e \in E_i)\right)$ *and* $\ln(\phi(e))$ *are convex, for every* $p(\cdot)$*.*

Proof. Turn the problem into a minimization by multiplying by minus one and take logarithm of the product. ■

By analogy with what has been showed for multi-dimensional q results can be derived for the Multi-Client case.

4.1 FOA Conditions for some common distributions

As it was seen in the previous section, conditions (16)-(17) can guarantee the validity of the first order approach in many different contexts. Therefore it is useful to study the distributions $f(x|e)$ that will satisfy them. One of the most used probability distributions in practice is the Normal distribution. It might seem difficult to verify condition (16) for a quality variable that is Normal for each e, since the Normal cumulative distribution function doesn't have a closed form. In the next result, conditions under which a Normal distribution that is affected by effort in its mean and variance, will satisfy (16), are derived.

Property 11 *Let* $F(q, \mu, \sigma)$ *be the cdf of the normal distribution with mean* $\mu(e)$ *and variance* $\sigma^2(e)$*. If* $R(\frac{y-\mu}{\sigma}) \cdot \sigma$ *is convex and non-increasing in effort for every* R *such that* $R' \geq 0$ *and* $R'' \geq 0$*, then* $F(x, \mu, \sigma)$ *will satisfy condition (16).*

Proof. See Appendix ■

A similar result can be derived for a truncated normal

Corollary 12 *Consider a r.v* q *with the following pdf,*

$$f_q(q|e) = \begin{cases} \frac{\frac{e^{-\frac{(t-\mu)^2}{\sigma}}}{\sigma\sqrt{2\pi}}}{d(e)} & \text{if } t \in [-c\sigma + \mu, c\sigma + \mu] \\ 0 & \sim \end{cases}$$

Where $d(e) = \int_{-r(e)-\mu}^{r(e)-\mu} \frac{e^{\frac{(x+\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} = \int_{-\frac{r(e)}{\sigma}}^{\frac{r(e)}{\sigma}} \frac{e^{\frac{x^2}{2}}}{\sqrt{2\pi}}$, *$\mu(e)$ and $\sigma(e)$ depend on e and c is a constant.*

If $R(\frac{y-\mu}{\sigma}) \cdot \sigma$ is convex and non-increasing in effort for every R such that $R' \geq 0$ and $R'' \geq 0$ and $c\sigma + \mu$ is concave, then $F(x, \mu, \sigma)$ will satisfy condition (16).

Corollary 13 *Let $F(q, \mu, \sigma)$ be the cdf of the log-normal distribution with mean $\mu(e)$ and variance $\sigma^2(e)$. If $R(\frac{y-\mu}{\sigma}) \cdot \sigma$ is convex and non-increasing in effort for every R such that $R' \geq 0$ and $R'' \geq 0$, then $F(x, \mu, \sigma)$ will satisfy condition (16).*

Similarly, the following result can be shown

Property 14 *If a random variable X that depends on a parameter θ, with cdf $F(x, \theta)$, has the property that through a change of variable we have that $F(x, \theta) = F(r(\theta)x, 1)$, then, if X depends on e only through θ and $R(y \cdot r(\theta))/r(\theta)$ is convex and non-increasing in effort for every R such that $R' \geq 0$ and $R'' \geq 0$, then*

- *$F(x, \theta)$ will satisfy condition (16).*
- *For any number of independent realizations of X, n, the distribution of the kth percentile as defined in (19), will satisfy condition (16).*

Property 15 *If a random variable X that depends on a parameter θ, with cdf $F(x, \theta)$, has the property that through a change of variable we have that $F(x, \theta) = F(r(\theta)x, 1)$, then, if X depends on e only through θ, then the non-Bernoulli utility function defined by (15) will be concave in e if (17) is satisfied.*

4.2 If First Order Approach conditions are not verified

Sometimes the conditions for the first order approach to be valid are not easily verifiable. This is the case, for example, when the utility function is a CARA, Property 10 will generally be difficult to verify and might not be true. Also, if the utility function used doesn't fulfill the separability conditions that we require in the previous sections, those properties will not apply. However, in many cases even if we don't have the sufficient conditions, FOA will still be valid as long as in the optimal contract the utility function of the provider has a unique stationary point in e given the payment schedule $p(\cdot)$. A heuristic would then be to pose the problem assuming that FOA is valid, find the corresponding payment schedule $p^*(\cdot)$ and effort e^* that maximize the providers utility and then verify the uniqueness of the stationary point in e ex-post, given the payment schedule $p^*(\cdot)$. In fact, even if there is not a unique stationary point one needs only to verify that the value of e^* is the global maximum of the utility of the provider, given $p^*(\cdot)$.

5. SLA determination in a n-tier IT Service Scenario

SLA/contracts for IT Services often contain clauses regarding desired levels of response time. To provide a certain level of response time, a service provider has to use costly resources. The response time obtained as a result will still be stochastic around an average value. In our framework an optimal menu of contracts has to take into account the randomness of any particular measure of response time, and the characteristics of the different types of clients. Consider a context in which to meet a prescribed metric of Response Time a Service Provider has to provision computing resources, in the form of compute servers. This is frequently the case for Application Service Providers such as e-commerce sites). In this article, we look at single tiered services (e.g. Database services, web server utilities, Application server utilities etc.) The Response Time of such a single-tiered IT Service is modelled using a simple analytic queuing theory model. We suppose we have an IT Service that receives requests, that arrive according to a Poisson process of parameter λ. Servers handle the requests and their service time will also behave as a r.v. If at any given time all servers are occupied with requests, all requests that arrive thereafter will wait in queue to be serviced. If we suppose that the requests, that one server receives, behave as a Poisson of parameter $\tilde{\lambda}$ and the service time of the server is exponential of parameter μ, it is a known result from queuing theory that the total Response Time will be exponentially distributed with parameter $\mu - \tilde{\lambda}$. If the workload is shared equally among the compute servers, then each server will receives requests at a rate λ/e, where e, the effort variable, is the number of servers, then the Response Time will distribute exponentially with parameter $\bar{\lambda}(e) = \mu - \frac{\lambda}{e}$.

SLA/contracts vary, in terms of the performance metric of response time that is used. The quality variable that is appropriate to use in the developed framework will depend on the particular performance metric in which any particular contract is written. For example, if a contract specifies, that the hourly average Response Time is lower or equal than 25 ms, the quality variable we propose for this case would be the realized hourly Response Time. This quality variable will follow a probability distribution that can be derived under this framework in which each Response Time behaves exponentially. In general, a befitting quality variable will be one that appears explicitly in the SLA to specify a determined level of service, and in terms of which penalties will depend. In the previous example, the clause could specify penalties such as, the SP will pay a penalty of $10,000$ if the average response time is between 25 and 35 ms and $20,000$ if the average response time is between 35 and 45 ms.

In this work we focus on some likely SLA clauses: (a) contracts that specify a desired average of response time lower than t, over a period of length $\bar{T}$; (b) contracts that are in terms of percentiles, such as 95% of the requests have a response time lower than t, over a period of length $\bar{T}$; (c) combinations of the two previous types, such as 95% of the hourly averages have to be lower or equal than t, over a period of length $\bar{T}$.

5.1 Mean Response Time

If an SLA is written in terms of the Average Response Time, computed during a length of time $\bar{T}$. If the total number of requests is n and the respective realized Response Times are $t_1, t_2, t_3, \ldots, t_n$, a possible Quality Variable, to fit this context, would be $-\bar{t}^n = -\frac{\sum_{i=1}^n t_i}{n}$, that is, minus the Average of Response Time [4]. The distribution of the average as an statistic depends on the original distribution of the sample.

5.1.1 Distribution of the Mean if Response Times are Exponentially Distributed

If we assume that the Response Times are exponentially distributed, the distribution of the mean Response Time, conditional on the total number of requests, n, will be a Gamma(n,$\frac{1}{n\bar{\lambda}}$)

$$f_{-\bar{t}^n}(t) = \frac{(\bar{\lambda}n)^n t^{n-1} e^{-\bar{\lambda}nt}}{\Gamma(n)} \tag{18}$$

If we know that the process of arrival of requests is also exponentially distributed, we can determine the distribution of the Average Response Time as

$$f_{-\bar{t}}(t) = \sum_{n=0}^{\infty} \frac{e^{-\lambda} e^{\bar{\lambda}nt} l^n \left(\bar{\lambda}n\right)^n t^{-1+n}}{\Gamma(n)\Gamma(1+n)}$$

However, for the sake of computational simplicity, we could estimate the quantity of requests by its mean, $\lambda\bar{T}$, in which case to determine the distribution of $\bar{t}$ we would have to replace $n = \lambda\bar{T}$ in (18).

This quality variable satisfies the conditions to use the FOA approach as it can be seen in Property 14.

5.1.2 Normally Distributed Mean

If the sample is big enough, from the Central Limit Theorem, we can assume that the mean has a Normal Distribution. This is useful if we are not sure of the underlying distribution of each response time or of any other quality variable that we are analyzing. In this case, since each response, has mean $\mu = \frac{1}{\lambda(e)}$ and variance $\sigma^2 = \frac{1}{\lambda(e)^2}$, a normally distributed mean will be a $N(\frac{1}{\lambda}, \frac{1}{n\lambda^2})$. However, assuming normality of response time might not be so appropriate for this particular case because the Normal distribution takes on any value in the real line, and response time is positive. An alternative would be to use a truncated normal distribution, that is for some positive function $r(e)$ with $-r(e) + \mu \geq 0$ we will have that

$$f_{-\bar{t}}(t) = \begin{cases} \frac{\frac{e^{-\frac{(t+\mu)^2}{\sigma}}}{\sigma\sqrt{2\pi}}}{d(e)} & \text{if } t \in [-r(e) - \mu, r(e) - \mu] \\ 0 & \sim \end{cases}$$

[4] Note that utility has to be increasing in quality

Where $d(e) = \int_{-r(e)-\mu}^{r(e)-\mu} \frac{e^{\frac{(x+\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} = \int_{-\frac{r(e)}{\sigma}}^{\frac{r(e)}{\sigma}} \frac{e^{\frac{x^2}{2}}}{\sqrt{2\pi}}$. Note that if $r(e) = constant \cdot \sigma(e)$, $d(e) = d$ will not depend on e. In that case $-\bar{t}$ will satisfy conditions (16)-(17) as it is shown in the appendix, using Property 11.

5.2 Percentiles

It is common that a Service Level Requirement is in terms of the quantile of response time, such as "95% of the realizations of Response Time have to be less than 5 seconds". A quality variable for this type of contract could then be related to the realized 95 percentile of response time, measured during a time period.

We will use the following definition of the kth percentile. If n is the number of requests, and $x_1, x_2, \ldots, x_n$ represent the ordered values of Response Time.

$$P_k = x_j \qquad \text{, where } j = round((n-1) \cdot \frac{k}{100} + 1) \tag{19}$$

Let T_k^n be the RV of the kth percentile given that there were n requests during the measurement period. Let's assume that the distribution of Response Times are iid and if $F(t)$ is the cumulative distribution function of each Response Time. The kth percentile is an order statistic, therefore its distribution will be,

$$f_{T_k^n}(t) = \frac{d}{dt} \sum_{i=j}^{n} I\!P(t_1 \leq t, t_2 \leq t, \ldots, t_i \leq t, t_{i+1} \geq t, \ldots, t_n \geq t) = \frac{n!}{(j-1)!(n-j)!} F(t)^{j-1} (1-F(t))^{n-j} f(t). \tag{20}$$

where j is given by (19).

5.2.1 Distribution of the Percentile if Response Times are Exponentially Distributed

If we assume that the Response Time of the servers system is exponentially distributed. In (20), we would have that $t_i \quad i \in \{1, 2, \ldots, n\}$ are r.v. iid, exponentially distributed, with parameter $\bar{\lambda} = \mu - \frac{\lambda}{e}$. This is not consistent with the fact that we are assuming that only n events took place, however if n is much bigger than the quantity of people in queue at any given time, the assumption becomes reasonable.

The probability distribution function of T_k^n will be:

$$f_{T_k^n}(t) = \frac{n!}{(j-1)!(n-j)!} (1 - e^{-\bar{\lambda} \cdot t})^{j-1} (e^{-\bar{\lambda} \cdot t})^{n-j} \bar{\lambda} e^{-\bar{\lambda} \cdot t}.$$

where j is defined by (19).

We will take n to be the mean of arrivals, that is $n = \lambda \bar{T}$, where $\bar{T}$ is the length of the measurement period.

If we take the quality variable to be minus the kth percentile of response time, we would have that the utility of the client is increasing in quality. Under this conditions, using Property 14 it can be verified that the FOA conditions (16)-(17) will be fulfilled.

5.3 Contracts that are in terms of means and percentiles

Using what has been discussed above, we can easily derive the probability distributions for quality variables that have to be in terms of percentiles and means.

1. If a contract is in terms of percentiles of averages, such as 95% of the hourly averages have to be lower or equal than t, over a period of length $\bar{T}$, its probability distribution will be given by (20), where F will be a Gamma or a Normal, depending on which distribution we choose to represent the mean. Later, using property 14, FOA conditions can be verified.

2. If a contract is in terms of averages of percentiles, such as the average of the hourly 95 percentiles have to be lower or equal than t, over a period of length $\bar{T}$, we know what the probability distribution of the percentiles will be and the distribution of the average of percentiles has to be determined. However, in this case, the Central Limit Theorem will tell us that if the average is taken over a big sample of percentiles, the average will be Normal, in which case we will need only to determine the mean and variance of the percentiles.

6. A Numerical Example

We computed optimal menus of contracts for different scenarios, using different utility functions and quality variables.

Two particular types of Utility functions were considered. Given a random variable X, that describes the behavior of uncertain profits, and with probability density function $f_X(x)$

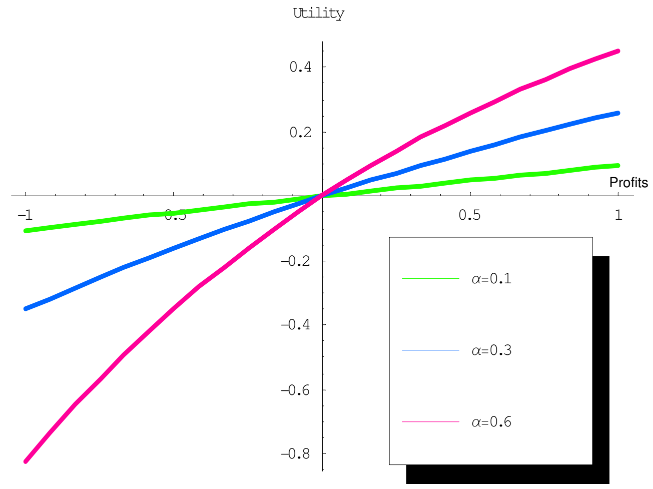
1. **CARA Utility Function**
   For outcome x, the utility will be $u(x) = (1 - e^{-\alpha x})$ the Expected Utility will then be $U(X) = \int u(x) f_X(x)dx$. The CARA Utility Function has constant risk aversion equal to α. In figure 3, it can be seen that the greater the parameter α the more concave the CARA function is and the more the agent dislikes risk. In our framework we will have $U(p(g), e) = u(p(g) - \phi(e))$ where $\phi(e)$ is convex in e.

2. **Exp-Var function** $U(X) = \mathbb{E}(X) - \tau \int (x - \mathbb{E}(X))^2 f_X(x)dx$
   The first term is the expected profits, τ is a positive constant and the second term is the variance of profits. The second term conveys the risk aversion of the agent.

6.1 Scenario: Mean Response Time and CARA Utility Function

We considered a case in which the quality variable is $q = -\bar{t}$, where $\bar{t}$ is the the realized Average Response Time, and is Gamma distributed. The g function, which represents the monetary valuation

Figure 3: CARA Utility Function



that a client gives to the quality variable, was taken to be of the following form

$$g_{k,m,t}(x) = \begin{cases} m(q-t)+k; & \text{if } x \leq t \\ k & \text{if } x \geq t \end{cases}$$

Different types of clients are parameterized by having different values k, m and t. The g function will increase linearly in $-\bar{t}$, with slope m, until a point in which it becomes constant and equal to k. The reason to assume this form of g function is that we assume that the clients value more quality up to a certain point in which they "saturate": greater quality does not increase his monetary utility any further. In this scenario we assume that the clients and the SP have CARA utility functions.

We will let the parameter of risk aversion α to vary for all agents. The value of $\bar{T}$ was taken to be 0.5 and we assume there is the same proportion of clients from each client type. The optimization problem was solved using the First Order Approach although conditions given in property 10 were not verified. The validity of the approach was verified ex-post. The set of efforts per client will be bounded below. For practical reasons in the numerical computations, for each client type θ_j, e_{θ_j} was taken such that $e_{\theta_j} \in [\frac{\lambda_{\theta_j}}{\mu_{\theta_j}} + \varepsilon, \infty)$, with ε small. We don't include the multipliers that correspond to the lower bounds of effort in the First Order Condition that represents (10), because we assume that the solution of (10) will be interior, which is later verified in practice.

6.1.1 Varying Risk Aversions

As a first analysis let's suppose that the SP is facing clients that have the same valuations of the service, that is, the same g function, that is shown in figure 4, but they differ in the risk aversion parameters α. Clients of type 1, 2 and 3 will have risk aversion equal to 3, 1.5 and 0.1, respectively. The values of $\bar{V}$ were varied with α keeping the certainty equivalent fixed. In tables 1 and 2 we present the parameters used for the computation of the optimal menu of contracts and in table 3 we present the optimal menu of contracts obtained.

Figure 4: "g function" (m=10, k=150000, t=-2)

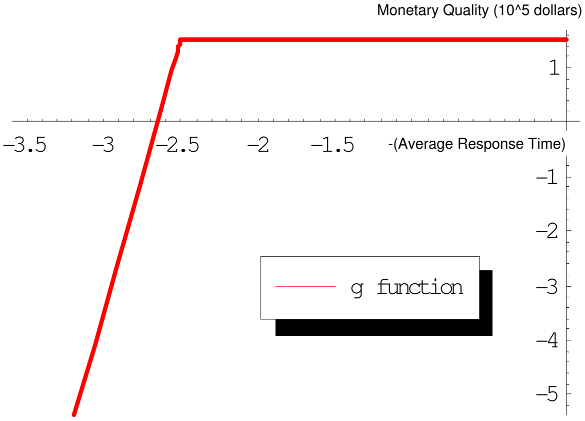


Table 1: Parameters Service Provider

Cost	Parameter of SP
2200	1.5

Table 2: Parameters Clients

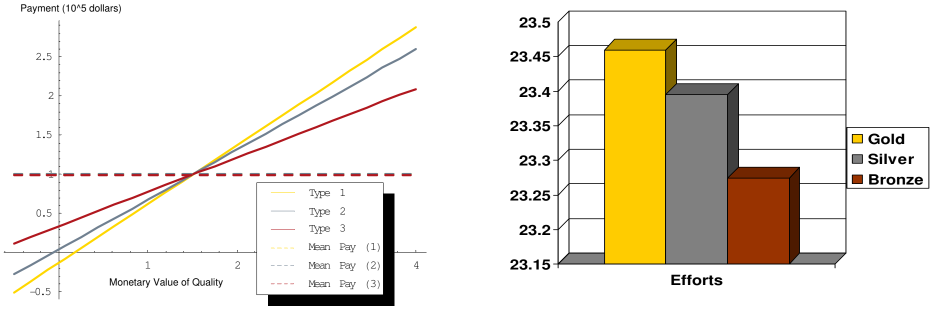
Type of Client	$\bar{V}$	mu	lambda	Risk Parameter of Clients
Type 1	0.78	5	100	3
Type 2	0.53	5	100	1.5
Type 3	0.049	5	100	0.1

In figure 5 we present a plot of the contracts obtained. The dashed lines represent the mean payments that each client type will make. The clients of type 1 who are the most risk averse will pay more for qualities that are above the mean than the other client types, and will pay less for qualities under the mean. The opposite is true for clients of type 3 who are the least risk averse.

Table 3: Optimal Menu of Linear Contracts

Type of Client	**Level of Effort**	**Slope**	**Intercept**
Type 1	23.46	0.75	-0.13
Type 2	23.39	0.64	0.04
Type 3	23.27	0.44	0.34

Figure 5: Different Risk Aversions



If now we change the risk aversion parameter of the service provider from 1.5 to 0.5, the slopes of the linear contracts become higher for all three client types, as it is shown in figure 6. In order to asses the optimality of the menu of contracts presented we pose ourselves the question of what the profits would be if the SP offered a different menu of contracts. If the SP offers only one contract to all clients, the self-selection constraints will be satisfied trivially. If we were to offer only one of the contracts of the three presented in table 3 it would have to be the contract offered to the clients of type 1, because the contract intended for their own type is the only contract that clients of type 1 are willing to accept. In figure 7 (left) we present the losses that the SP would experience if she offered such contract, with respect to the optimal menu of contracts in table 3. Profits made from payments of clients of type 1 will remain constant, while profits from the other two client types will be lower. In table 4 we present the optimal contract that the SP would offer to each client type if she could know which type is which, we refer to this as the "perfect discrimination case". If the SP were to offer one of the three, as before, the only contract that would be accepted by type 1 clients would be the one that is optimal for their client type. In figure 7 (right) we present the differences between this contract and the optimal menu of contracts in table 3. The losses are lower in this case, but they are positive. Note also that these client types are identical in every respect except for their risk aversion factors.

Figure 6: Changing Risk Parameter of Service Provider

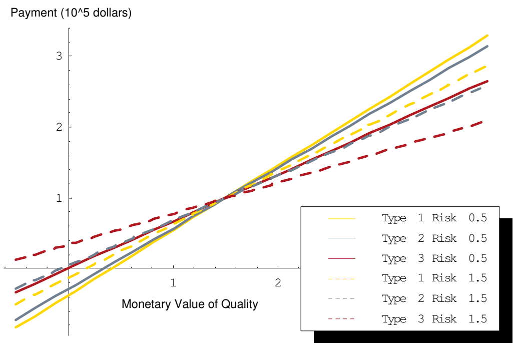
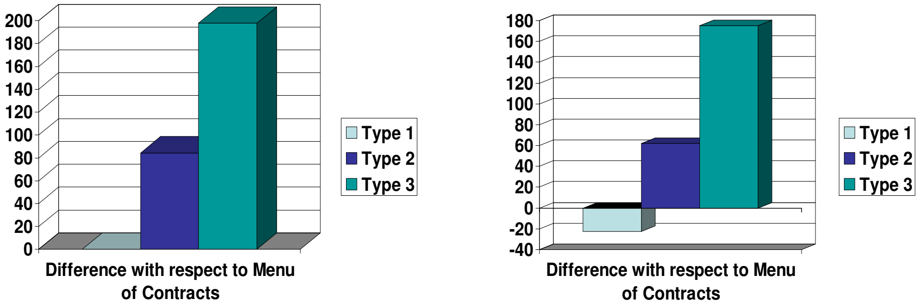


Table 4: Optimal Contracts with Perfect Discrimination

Type of Client	**Level of Effort**	**Slope**	**Intercept**	**Profits per Client Type**
Type 1	23.44	0.70	-0.05	48350.25
Type 2	23.37	0.59	0.12	48443.54
Type 3	23.28	0.44	0.34	48573.78

Figure 7: Comparisons with Optimal Menu of Contracts

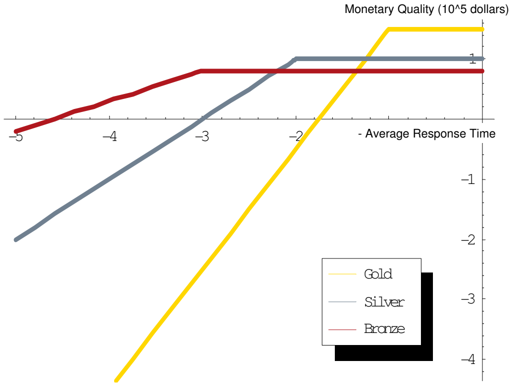


6.1.2 Gold, Bronze and Silver Clients

The framework presented above allows us to compute optimal menu of contracts for clients who were different in many dimensions. In figure 8 we present the g function for three hypothetical types of clients, the first type will have parameters $m = 1, k = 1, t = -2$, the second, $m = 2, k = \frac{3}{2}, t = -1$, and the third, $m = \frac{1}{2}, k = 0.8, t = -3$. We will refer to them as type "Gold", "Silver" and "Bronze", respectively, and we will assume that there is the same amount of clients of each type.

The second type values high quality of service more than the other two (higher k), however, his profits decrease faster as quality goes down, also his saturation point is higher. The third type requires lower levels of service, he has a low saturations point, and his profits don't decrease very fast if quality becomes lower, he also values the highest quality less. The first one would be the "middle" type. We will give each client type a different parameter α and a different value of $\bar{V}$.

Figure 8: Gold, Silver and Bronze Clients



In tables 5 and 6 we present the parameters for the SP and for each client type. The Gold clients will have a higher risk aversion parameter and $\bar{V}$, and the Bronze clients will have the lowest value for those two parameters.

Table 5: Parameters Service Provider

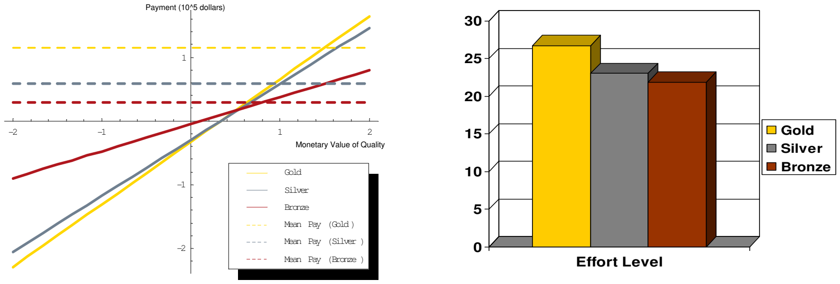
Cost	Parameter of SP
1000	0.1

In figure 9 we present the optimal linear contract and the respective levels of effort and in dashed lines, the mean payments. The SP will make more more profits from clients of type Gold, as it can be seen in figure 10.

Table 6: Parameters Clients

Type of Client	$\bar{V}$	**mu**	**lambda**	**Risk Parameter of Clients**
Gold	0.1	5	100	0.3
Silver	0.08	5	100	0.2
Bronze	0.05	5	100	0.1

Figure 9: Optimal Contract



The contracts have to be translated into (true) quality. In figure 11 we present the contracts in terms of (true) quality, and in dashed lines the mean payments for each client type. Note that the mean payments will be very close to the highest payments possible. This is because the efforts assigned to each client, for this example, will deliver a quality inside the area of saturation with high probability.

Figure 10: Profits from each client type

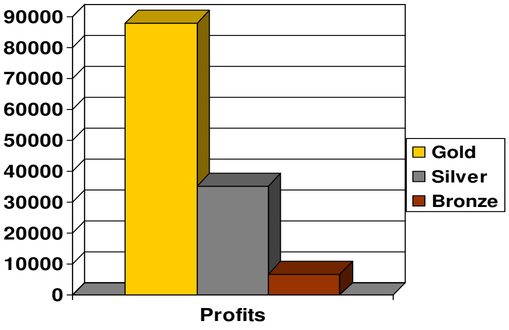
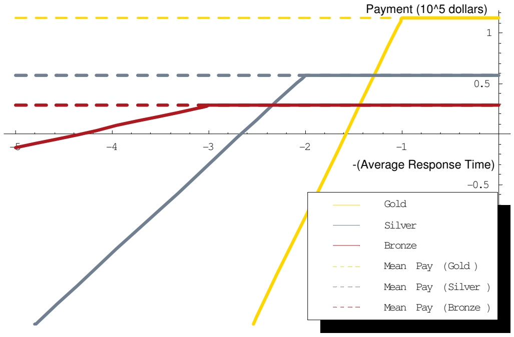


Figure 11: Quality vs Payment



6.2 Scenario: Mean Response Time and Exp-Var Utility Function

Very similar results are obtained when the Exp-Var utility function is used. For this function we didn't verify sufficient conditions for the FOA approach, but its validity is confirmed ex-post. The value of $\bar{T}$ was taken to be 2 and we assume there is the same proportion of clients from each client type. The "g" function that was used is the same as in the first part of the previous scenario, and is showed in figure 4. In tables 7 and 8 and in figure 12 we represent graphical representations of the contracts obtained.

Table 7: Parameters Service Provider

Cost	**Parameter of SP**
2200	6

Table 8: Parameters Clients

Type of Client	$\bar{V}$	**mu**	**lambda**	**Risk Parameter of Clients**
Type 1	0.5	5	100	3
Type 2	0.5	5	100	1.5
Type 3	0.5	5	100	0.1

Table 9: Optimal Contracts

Type of Client	**Level of Effort**	**Slope**	**Intercept**
Type 1	23.02	5.95	-7.93
Type 2	22.99	4.78	-6.17
Type 3	22.89	1.66	-1.49

Note that it is not straightforward to make comparisons between the contracts obtained using the CARA utility function and the Exp-Var. If the agents have different utility functions, their preferences will be different, even if both utility functions were appropriate to represent their preferences, a calibration of the parameters has to be made. However, we can see from the results obtained (see table 9 and figure 12) that we get analogous interpretations for the contracts for different clients given their risk parameter.

If now we decrease the risk parameter of SP from 6 to 2, we can see in figure 13 that the slopes of the contracts for every client will increase, this has a similar interpretation as in the previous case.

Figure 12: Optimal Contract

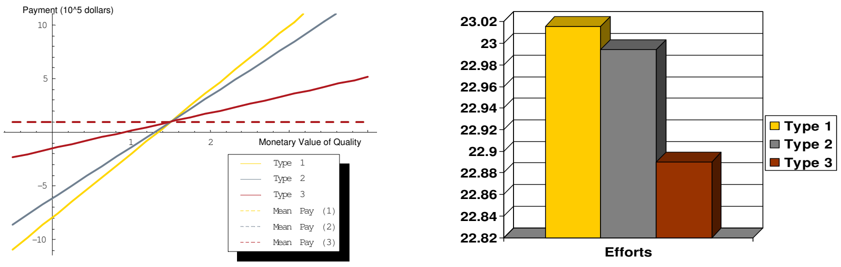
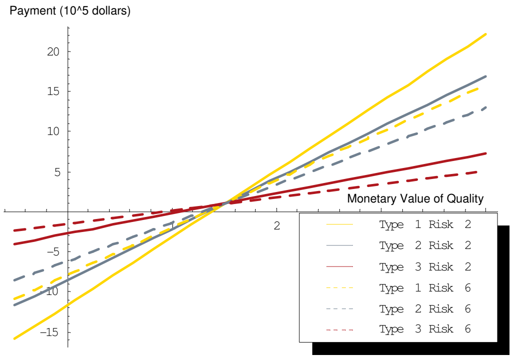


Figure 13: Optimal Contract



7. Conclusions

In this paper we propose a game theoretic approach to find an optimal SLA/Contract given the characteristics of the client and service provider. We extend the first basic model to take into account the possibility that the service provider will offer different contracts to different client types, in order to cover different necessities and extract more profits. We presented also a numerical example, in which a number of insights, that were consistent with economical intuition, were obtained. We also analyzed the conditions under which a First order Approach can be used, based on the literature of the principal-agent problem. In practice, the usefulness of our model, if calibrated correctly, is to give benchmarks for future contracts, in each stage of an eventual negotiation process. For calibrations purposes, information of past contracts can be used. It is important to note that there is still are aspects of the determination of SLA/Contracts that are not tackled here. For instance, our framework requires that the agents have a great deal information about each other. This is not generally the case, since, in reality, there might exist a gap of information between the agents. A possible line for future research that could be accounted for, and in turn, this could also give us a greater understanding on how the negotiation process takes place. Also, research on the behaviour of IT systems, given the amounts of inputs that are used, is of great importantance in this context, since it will affect the nature of the contract to be signed.

Acknowledgements

We would like to thank Rich Friedrich, Sharad Singhal, John Wilkes, Pano Santos, Alex Zhang, Lola Bautista and Cristian Figueroa for their helpful comments.

References

[1] George Candea and Armando Fox, 2002, A Utility-Centered Approach to Building Dependable Infrastructure Services *Proceedings of the 10th ACM SIGOPS European Workshop (EW-2002)*, pp. 213-218, Saint-milion, France, September 2002.

[2] Haluk Demirkan, Michael Goul, Daniel S. Soper, 2005, Service Level Agreement Negotiation: A Theory-based Exploratory Study as a Starting Point for Identifying Negotiation Support System Requirements, *Proceedings of the 38th Hawaii International Conference on System Sciences - 2005.*

[3] Jewitt, Ian, 1988, Justifying the First-Order Approach to Principal-Agent Problems, *Econometrica 56* ,(5), 1117-1190.

[4] R. Tyrrell Rockafellar, Stanislav Uryasev, Michael Zabarankin, 2002, Deviation Measures In Risk Analysis And Optimization, *RESEARCH REPORT # 2002-7* Risk Management and Financial Engineering Lab Center for Applied Optimization Department of Industrial and Systems Engineering University of Florida, Gainesville.

[5] Akhil Sahai, Anna Durante, Vijay Machiraju, 2002, *Software Technology Laboratory, HP Laboratories*, Palo Alto, HPL-2001-310 (R.1).

[6] Stole, Lars, 2001, Lectures on the Theory of Contracts and Organizations.

8. Appendix

8.1 Multi-Client Model

Consider the following forms of utility function of the Service Provider,

$$U(\{p_{\theta_j}(g_{\theta_j}(q^{i,\theta_j}))\}_{i,j}, \sum_{j,i} e_{i,\theta_j}) = \sum_{j,i} \phi_j(p_{\theta_j}(g_{\theta_j}(q^{i,\theta_j})) - \Phi(\sum_{j,i} e_{i,\theta_j}) \tag{21}$$

or

$$U(\{p_{\theta_j}(g_{\theta_j}(q^{i,\theta_j}))\}_{i,j}, \sum_{j,i} e_{i,\theta_j}) = \left(\prod_{j,i} \phi_j(p_{\theta_j}(g_{\theta_j}(q^{i,\theta_j}))\right) \Phi(\sum_{j,i} e_{i,\theta_j}) \tag{22}$$

with $\frac{\partial U}{\partial p_{\theta_j}(g_{\theta_j}(q^{i,\theta_j}))} \geq 0$, $\frac{\partial U}{\partial e_{i,\theta_j}} > 0$ $\frac{\partial^2 U}{\partial^2 p_{\theta_j}(g_{\theta_j}(q^{i,\theta_j}))} \geq 0$, $\frac{\partial^2 U}{\partial^2 e_{i,\theta_j}}' \geq 0$.
The optimization problem will be much simplified. In the first case we can rewrite the objective function as,

$$\int U(\{p_{\theta_j}(g_{\theta_j}(q^{i,\theta_j}))\}_{i,j}, \sum_{j,i} e_{i,\theta_j}) \prod_{j,i} f_{\theta_j}(q^{i,\theta_j}|e_{i,\theta_j}) = \sum_{i,j} \int \phi_j(p_{\theta_j}(g_{\theta_j}(q))) f_{\theta_j}(q|e_{i,\theta_j}) dq - \Phi(\sum_{i,j} e_{i,\theta_j}) \tag{23}$$

In the second case,

$$\ln\left(\int U(\{p_{\theta_j}(g_{\theta_j}(q^{i,\theta_j})\}_{i,j}, \sum_{i,j} e_{i,\theta_j}) \prod_{j,i} f_{\theta_j}(q^{i,\theta_j}|e_{i,\theta_j})\right) = \sum_{i,j} \ln(\int \phi_j(p_{\theta_j}(g_{\theta_j}(q))) f_{\theta_j}(q|e_{i,\theta_j}) dq) - \ln(\Phi(\sum_{i,j} e_{i,\theta_j})) \tag{24}$$

If the U fulfills properties (21) or (22) there will not be unicity in (10). In particular, if there is a solution $\{e_{i,\theta_j}\}_{i\in\{1,\ldots\mu_i\}, j\in\{1\ldots k\}}$, any other set of efforts $\{\bar{e}_{i,\theta_j}\}_{i\in\{1,\ldots\mu_j\}, j\in\{1\ldots k\}}$ such that $\sum_{i=1}^{\mu_j} e_{i,\theta_j} = \sum_{i=1}^{\mu_j} \bar{e}_{i,\theta_j} \quad \forall\theta_j$ will also be a solution to (10). We will assume without loss of generality that the solution will be symmetric. That is, $e_{r,\theta_j} = e_{s,\theta_j} = e_{\theta_j} \quad \forall s, t$, and the maximization problem in (7)-(10) will be over $\{p_{\theta_j}(\cdot)\}_j, \{e_{\theta_j}\}_j$. In such case (23) will become,

$$\sum_j \mu_j \int \phi(p_{\theta_j}(g_{\theta_j}(q))) f_{\theta_j}(q|e_{\theta_j}) dq - \Phi(\sum_j \mu_j e_{\theta_j})$$

and (24) will be,

$$\sum_j \mu_j \ln\left(\int \phi_j(p_{\theta_j}(g_{\theta_j}(q))) f_{\theta_j}(q|e_{\theta_j}) dq\right) - \ln\left(\Phi(\sum_j \mu_j e_{\theta_j})\right)$$

8.2 Foa Approach

Proof of Property (6). Let $[\underline{q}, \bar{q}]$ be the support of the quality q, and let $\tilde{u} : [\underline{q}, \bar{q}] \longrightarrow I\!R$ be a concave and non-decreasing function,

$$\int_{\underline{q}}^{\bar{q}} \tilde{u}(q) f(q|e) dq = \tilde{u}(\bar{q}) F(\bar{q}|e) - \tilde{u}'(\bar{q}) \int_{\underline{q}}^{\bar{q}} F(q|e) dq$$

$$+ \int_{\underline{q}}^{\bar{q}} \tilde{u}''(\tilde{q}) \left(\int_{\underline{q}}^{\tilde{q}} F(q|e) dq \right) d\tilde{q}$$

$F(\bar{q}|e) = 1$ and $\int_{\underline{q}}^{\bar{q}} F(q|e)dq = \bar{q}F(\bar{q}|e) - \int_{\underline{q}}^{\bar{q}} qf(q|e)dq$ therefore, if (16)-(17) are satisfied, differentiating inside the integral we can see that the first expression will be concave in e. Now, we conclude notint that if p is concave and non-decreasing, then $u(p(g(q)))$ will be concave and non-decreasing in q. ■

8.3 Common Distributions

Proof of Property 11.

$$\int_{-\infty}^{y} F(x, \mu, \sigma) dx = \int_{-\infty}^{y} \Phi(\frac{x - \mu}{\sigma}) dx = \int_{-\infty}^{\frac{y-\mu}{\sigma}} \Phi(x) dx \sigma$$

The function $R(t) = \int_{-\infty}^{t} \Phi(x)dx$ is convex and increasing in t since its first derivative is a cumulative function and the second derivative a probability distribution function. Therefore, to know if for a particular example will satisfy (16) it is only necessary to prove that $R(\frac{y-\mu}{\sigma}) \cdot \sigma$ is convex and nonincreasing in effort, for convex and increasing R. ■

8.4 Example

Proof of FOA for Mean Response Time. We want to prove that (16) is satisfied.

$$F_{-\bar{t}}(t) = I\!P(-\bar{t} \leq t) = I\!P(\bar{t} \geq -t) = \begin{cases} \frac{Gamma(n, n\bar{\lambda}(-t)))}{\Gamma(n)} & \text{if } t \leq 0 \\ 1 & \sim \end{cases}$$

Now,

$$\int_{-\infty}^{y} F_{-\bar{t}}(t) dt = \begin{cases} -\frac{-y\Gamma(n, n\bar{\lambda}(-y))}{\Gamma(n)} + \frac{\Gamma(1+n, n\bar{\lambda}(-y))}{n\bar{\lambda}\Gamma(n)} & \text{if } y \leq 0 \\ \frac{1}{\lambda} + y & \text{if } \sim \end{cases}$$

$\frac{1}{\lambda}$ is nonincreasing and convex with respect to e, we need only to analyze the case $t \leq 0$.

$$\frac{d}{de} \int_{-\infty}^{y} F_{-\bar{t}}(t) dt = -\frac{\lambda\Gamma(1 + n, \left(-\frac{\lambda}{e} + \mu\right) n(-y))}{(\lambda - e\mu)^2 \Gamma(1 + n)} \tag{25}$$

$$\frac{d^2}{d^2 e} \int_{-\infty}^{y} F_{-\bar{t}}(t) dt = \frac{\lambda(-y) \left(\left(-\frac{\lambda}{e} + \mu\right) n(-y) \right)^n \left(e^{\frac{(\lambda - e\mu)n(-y)}{e}} \lambda + 2e\mu \text{ExpIntegralE} \left[-n, \left(-\frac{\lambda}{e} + \mu\right) n(-y) \right] \right)}{e^2 (\lambda - e\mu)^2 \Gamma(n)} \tag{26}$$

We conclude noting that (25) and (26) are negative and positive, respectively.[5] ■

Proof of (16) for Normal Mean Response Time. Let's define $\mu = \frac{1}{\lambda}$ and $\sigma = \frac{1}{\lambda}$, let $-r(e) + \mu \geq 0$ and $r(e) = c\sigma(e)$ where $c \leq 0$ is a constant.

$$\int_{infty}^{y} F_{-\bar{t}}(t)dt = \begin{cases} 0 & \text{if } y \leq -r(e) - \mu \\ R(\frac{(y+\mu)n}{\sigma})\frac{\sigma}{d \cdot n} & \text{if } y \in [-r(e) + \mu, r(e) + \mu] \\ R(\frac{r(e)}{\sigma})\sigma + (y - r(e) + \mu) & \sim \end{cases}$$

For the middle case, since $\frac{R(n+nx\bar{\lambda}(e))}{\bar{\lambda}(e)n}$ is convex and nonincreasing in e, where λ. The second derivative with respect to e is

$$\tfrac{1}{\bar{\lambda}(e)^3}\left(\left(R(n + nx\bar{\lambda}(e)) - nx\bar{\lambda}(e)R'(n + nx\bar{\lambda}(e))\right)\left(2\bar{\lambda}'(e)^2 - \bar{\lambda}(e)\bar{\lambda}''(e)\right) + n^2x^2\bar{\lambda}(e)^2\bar{\lambda}'(e)^2R''(n + nx\bar{\lambda}(e))\right)$$

which is positive, since $\bar{\lambda}'(e) \geq 0$, $\bar{\lambda}''(e) \leq 0$ and x is negative.

We conclude noting that $-r(e) + \mu = c_0\frac{1}{\bar{\lambda}(e)}$, with c_0 is a positive constant, and that $\frac{1}{\bar{\lambda}(e)}$ is convex. ■

[5] $\Gamma[a, b] = \int_b^\infty t^{a-1}e^{-t}dt$ is the Incomplete Gamma Function, and ExpIntegralE$[n, z] = \int_1^\infty \frac{e^{-zt}}{t^n}dt$ is the exponential integral function.