OPTIMAL INFORMATION TECHNOLOGY SERVICE PRICING AND CAPACITY DECISION UNDER SERVICE-LEVEL AGREEMENT: A PARAMETRIC ANALYSIS

Research-in-Progress

Xiao Zou
Department of Information Systems
National University of Singapore, Singapore
zouxiao@comp.nus.edu.sg

Ke-Wei Huang
Department of Information Systems
National University of Singapore, Singapore
huangkw@comp.nus.edu.sg

Abstract

Service-level agreements (SLA) are essential for IT service management due to the complexity of IT service delivery model and the importance of ensured Quality of Service (QoS) for corporate clients. This study proposes a novel geometric framework for analyzing the optimal pricing and capacity planning for IT service provider under SLA. The IT service with congestion costs is modeled using an M/M/c queueing model. Using analytical modeling approach, our preliminary analysis shows that the optimal pricing and capacity depend critically on the functional form of the SLA contract. Moreover, the comparative statics shows how the optimal prices are affected in response to the changes in parameters defined in SLA. Our study contributes to the theoretical understanding of IT service management by introducing an innovative approach to visualize the revenue-SLA trade-offs. This study also helps service providers and business clients enhance their understanding about price/penalty scheme during contract negotiation or service-level management.

Keywords: Pricing, Parametric curve analysis, Service-level agreement
Introduction

Services-orientated business model in IT industry is one of the fastest-growing paradigms in the global economy (Bardhan et al. 2010). According to Gartner (2013), worldwide IT services market is expected to reach over one trillion by 2016, growing at a 4.5% average compound annual growth rate through 2016. Due to the complexity of service delivery technical model, the performance of service providers may not meet the expectation of the corporate clients. As a result, service level agreement (SLA) is often specified to define the Quality of Service (QoS) in the IT service contract. In practice, SLAs of IT service are typically defined by using technical terms such as mean or quantiles of response time, annual uptime percentage, throughput, jitter and other performance metrics. For example, Amazon's EC2 web service is guaranteed to be available with an annual uptime percentage of at least 99.95% during the Service Year.¹ In AT&T's business network service's SLA, the monthly average latency is guaranteed to be less than 40.0 ms. And customers may get refund for one day's credit out of monthly bill, if AT&T fails to meet the SLA.² However, pre-specification of price and quality levels in SLA poses significant management challenges for IT service provider because of the demand uncertainty from clients' service usage (Sen et al. 2009).

Most SLAs often include a penalty scheme when the service provider does not meet the pre-specified service quality threshold. Due to the uncertainty of the service demand at a specific time point, it may not be profit-maximizing for the service provider to invest huge amount of capital in expanding the capacity of the underlying infrastructure to meet the peak demand. In other words, IT service provider typically faces the trade-off between the revenue earned from delivering IT services and the potential loss due to the SLA violation (Sandmo 1971). To the best of our knowledge, there exists scarce literature that investigates how the functional form of SLA penalty scheme may affect the optimal pricing and capacity planning. Particularly, different firms may have to bear different intangible costs when breaching the SLA. For example, a large company may care more about the brand image so that breaching SLA is more costly for larger and reputable service providers.

In order to address the revenue and SLA penalty trade-off, this study considers a monopoly model with the objective function comprised of two key components: a standard revenue function with congestion costs (Mendelson 1985) and a penalty scheme if the terms of a SLA are broken. To maximize the objective function, the IT service provider can decide two decision variables: the usage-based price and the number of servers in an M/M/c queueing model. We first analyze the short-term optimal pricing problem given any number of c. Next, we extend the short-term model to address the long-term capacity planning problem: what is the profit-maximizing number of capacity c? Therefore, this research-in-progress aims to investigate the following research questions: (1) Given the long-term fixed service capacity, under different types of SLA penalty schemes, how does the short-term optimal pricing differ from the revenue-maximizing level without SLA penalty? (2) How does the optimal pricing change with respect to changes in SLA parameters under different SLA penalty schemes? (3) How to decide the long-term optimal capacity decision of IT service providers?

In this analytical modeling study, the IT service with congestion cost is modeled using an M/M/c queueing model, which is a multi-server queueing model that assumes Poisson arrival process of requests and each request requires exponentially distributed service time. This type of queueing model has been widely applied to investigate various IT services such as network services (Gupta et al. 2011), organizational databases (Konana et al. 2000), application service providers (Cheng and Koehler 2003), e-sourcing (Dai et al. 2005), Information networks (Chen et al. 2011) and websites performance (Liu et al. 2010). Queueing model allows us to model variations in response time, which is an important aspect of QoS in IT service industries.

We propose a novel geometric framework for analyzing the optimal pricing and capacity planning under various SLA formulations. The uniqueness of our model is to depict two components of the objective function on the two-dimensional plane as a parametric curve. On this plane, X-axis is the SLA penalty and Y-axis is the standard monopoly revenue function with congestion costs. In this way, we can visualize and analyze the properties of how one decision variable may affect the trade-off between two components in the objective function.

¹ Amazon EC2 Service Level Agreement: [http://aws.amazon.com/ec2-sla/](http://aws.amazon.com/ec2-sla/)
This study aims to make valuable contributions to the understanding of IT service management in several ways. First, the new parametric curve approach visualizes the trade-off between the ordinary revenue-maximization and the QoS management problem. This parametric analysis is simpler than other models proposed in the literature. We show that as long as a normalized ratio (similar to Sharpe Ratio in CAPM in Finance) between the SLA penalty \((X)\) and the standard revenue function \((Y)\) is decreasing (or increasing) in the decision variable, the usage price in this study, a number of neat results can be shown. Fortunately, we can show that under fairly general assumptions, this mathematical condition can be met. Second, this study explicitly models various SLA penalty scheme and focuses on the how SLA characteristics may affect the optimal usage pricing. In this way, our model contributes to IT service pricing literature by incorporating SLA into service provider’s objective function, which has not examined in the literature. Previous studies either neglected SLA in objective function (Dewan and Mendelson 1990), or treated price as exogenous variables when investigating SLA (Sen et al. 2009; Sen et al. 2010). Our model combines these two perspectives and attempts to derive the optimal pricing with the presence of SLA. Specifically, we examine three penalty metrics: fixed per-unit penalty, per-unit penalty based on price, or a proportion of total bill as lump-sum penalty if threshold Quality of Service (QoS) is breached. Lastly, this study also provides insights for practitioners to understand how the optimal capacity planning may be influenced by SLA’s penalty metrics.

**Literature Review**

This study is directly related to congestion pricing literature, which provides a model of a service system where the customers requesting service are rational, utility-maximizing, and delay-sensitive decision makers (Ata and Shneorson 2006). Specifically, Naor (1969) provided a classical analysis of resource allocation and pricing when congestion determines service quality. A seminal work by Mendelson (1985) incorporated a microeconomics approach to model the pricing for computing services. Following Mendelson (1985), Dewan and Mendelson (1990) extended the model by considering nonlinear delay cost function. Westland (1992) examined the demand price elasticity and derived the profit-maximizing prices in a short-run setting. Afeche and Mendelson (2004) proposed alternative price-service mechanisms for a provider that serves customers whose delay cost depends on their service valuations. Lastly, Ata and Shneorson (2006) studied a service facility in which the system manager dynamically controlled the arrival and service rate to maximize the long-run average value generated. Compared with the models in the literature, our model is unique as it incorporates the SLA penalty with optimal pricing and capacity decision. Unlike Westland (1992), our model examines the two-stage setting (short-term pricing and long-term capacity) and investigates the two distinct decisions separately. Moreover, the objective function for most of the abovementioned studies is the aggregated benefits of using the target IT service, not a standard profit function for monopoly pricing in the Industrial Organization literature. The demand structure in our model follows the conventional monopoly pricing setup and resembles that in Chen and Frank (2004), which derived a steady-state equilibrium demand function and proved several useful properties associated with the general demand structure.

There are a few recent studies examining the SLA in the context of IT services. In particular, Sen et al. (2009) proposed a dynamic price-penalty mechanism for SLA to capture demand uncertainty using simulation approach. Sen et al. (2010) explored the value of sharing demand information as a resource allocation mechanism for IT services. Moreover, Maglaras and Zeevi (2005) considered two levels of services, guaranteed and best-effort, and examined the performance-contingent pricing of these two types of queueing models for broadband network services. Our study examines a different property of SLA; we focus on how various types of penalty schemes may affect the optimal trade-off between standard revenue-maximizing in economics and loss from breaching SLA.

This study also relates to strategic analysis of long-term capacity problem of information systems. Gupta et al. (2011) studied the network capacity investment problem under congestion-based pricing and flat-rate pricing. They developed a heuristic to compute the maximum capacity a network provider would be willing to invest. Stidham (1992) and Dewan (1996) investigated the problem of service facility expansion under a variety of pricing and control structures using queueing framework. Huang and Sundararajan (2011) presented a model of usage pricing for digital goods with discontinuous infrastructure costs using non-linear pricing model in the literature. Our proposed model takes a similar perspective by considering the service capacity as a long-term decision variable under queueing framework.
Research Model

**IT Service Demand with Congestion Cost**

We model the firm’s core IT service using a pricing model with queueing effects pioneered by Mendelson (1985), which could be the most widely applied service-oriented pricing model. Queueing approach allows us to explicitly model waiting time, which is an important aspect of the service quality in service industries including cloud computing services, consulting and law firms and other service sectors. Consistent with the notation in the queueing literature, we use $\lambda$ to denote the service demand rate. In our model, $\lambda = D(p)$ and thus $\lambda$ is a decreasing function of $p$. In other words, when the IT service provider increases the usage price (i.e. pay-per-use) $p$, the requests rate will decrease. Our model assumes that in the equilibrium, the average waiting time must equal the expected waiting time $w$. Formally, in standard queueing model, the steady-state equilibrium condition is given by

$$V'(\lambda) = p + v \times w,$$

where $V(\lambda)$ is the value function that represents the gross value corresponding to demand rate $\lambda$. $v$ represents the marginal delay cost of a service request. The LHS of Equation (1) denotes the expected marginal value and the RHS denotes the marginal cost (price and congestion cost) of the service request.

To model the congestion effect, we further consider expected waiting time $w$ as a function of $p$ and service rate $\mu$, i.e. $w = w(D(p), \mu)$. In this way, the price has both direct effect (via pricing) and indirect effect (via delay cost) on service demand: higher $p$ directly reduces demand, but at the same time the reduced demand can relax congestion, reduce $w$ and increase demand. Chen and Frank (2004) have derived a steady-state Nash equilibrium of such demand. The solution of $\lambda$ in eq.(1) characterizes the equilibrium service demand, which we denote as $D(p)$ in this study. Following their results, $D(p)$ is a concave and decreasing function of $p$, i.e. $D'(p) \leq 0$, $D''(p) \leq 0$.

**Service Provider’s Objective Function**

There are two key components in the objective function of an IT service provider, who cares both the revenue from service provision and the expected loss due to SLA. Formally, the objective is given by

$$\Pi(p,c) = R(p) - \theta S(p,c),$$

where $R(p)$ is the revenue function of the service and is equal to $p \times D(p)$ throughout the paper. $S(p,c)$ is the expected loss due to SLA violation and depends on price $p$ and computing capacity $c$. IT service providers generally prefer higher revenue and lower SLA loss. $\theta$ denotes the weight of SLA loss in the objective function and it represents the importance of SLA for service provider. For example, when $\theta=1$, the service provider only cares about the monetary value of $S(p,c)$. When $\theta>1$, the service provider cares more about the SLA loss because violating SLA may also damage brand image. We assume zero marginal cost of service provision. Our results can be easily generalized to a setup with constant variable cost.

**Service Level Agreement Formulation**

This study considers three SLA formulations on the response time of IT service. In practice, a typical formulation is the per-unit penalty imposed to each breach event of SLA. For example, SLA could guarantee the response time within 500 milliseconds and a fixed amount of penalty will be fined for each violation. Formally, this SLA penalty function is:

$$S(p,c) = l \times \Pr(w > k) \times D(p),$$

where $w$ is the waiting time, $l$ is the per-unit penalty for each unit time delay, and $k$ is the threshold value of waiting time guaranteed in SLA. Both $l$ and $k$ are specified in the SLA contract. Hence, $\Pr(w > k)$ represents the probability of SLA violation and $\Pr(w > k) \times D(p)$ is the total number of violations.

A similar SLA formulation is to replace the fixed per-unit penalty by a proportion of the usage price. In
this case, the per-unit penalty depends on the usage price imposed on the service.

\[ S(p, c) = \lambda \times p \times \Pr(w > k) \times D(p), \]  

where \( \lambda \times p \) is now the per-unit penalty that depends on \( p \). This penalty scheme is also common in practice when the penalty is specified as a percentage of the usage price.

Moreover, many IT service providers adopt SLA scheme based on averaged waiting time during a sample period rather than on one violation instance. For example, AT&T’s business network service would refund customers one day credit of the monthly bill, if the monthly average latency is longer than 40ms. In this case the SLA penalty function is formalized as follows:

\[ S(p, c) = \begin{cases} \lambda \times p \times D(p) & \text{if } \bar{w} > k \\ 0 & \text{if } \bar{w} \leq k \end{cases}, \]  

where \( \bar{w} \) is the average waiting time during the sample period and \( k \) is the threshold QoS. The SLA loss is on the condition of \( \bar{w} \) and \( k \), both of which are defined in SLA penalty scheme. In this RIP, for simplicity, we only consider the asymptotic case: \( \bar{w} \) is a constant rather than a random variable. Details of the general case and other SLA schemes will be explored in future works.

**Short-term pricing Problem of IT service**

In line with standard firm decision problems in microeconomics, IT service provider generally faces two decision problems. In the short-run, given the long-term IT capacity (e.g. the number of servers \( c \)), the IT service provider needs to decide an optimal pricing \( p \) that maximizes the objective function defined in eq.(2). In the long-run, the IT service provider faces an information systems acquisition problem and needs to decide the optimal computing capacity \( c \) (Westland 1992) that maximizes objective function. Given any IT capacity, the short-term decision problem is given by

\[ \max_p \Pi(p, c) \tag{6} \]

For ease of exposition, we define the following two solution variables. The set of overall profit-maximizing solution is denoted by \( P^* \) and each optimal solution is denoted as \( p^* \) because it is possible that there are multiple solutions in general. As a result

\[ P^* = \arg \max_p \Pi(p, c) \tag{7} \]

By the results of Chen and Frank (2004), we know that the revenue function \( R(p) \) is unimodal and there exists a unique revenue-maximizing price \( p^{**} \) as follows

\[ p^{**} = \arg \max_p R(p) \tag{8} \]

In our framework, we use \( p^{**} \) as the benchmarking case and one goal is to compare \( p^{**} \) and \( p^* \). While extant studies in the literature typically only analyze \( p^{**} \) and ignore the loss due to SLA, we examine how firms (i.e. IT service providers) should adjust their pricing relative to \( p^{**} \) under different SLA penalty functions.

**Long-term Capacity Problem of IT service**

One major benefit of the software-as-a-service model is the centralized and shared IT infrastructure at the service provider (Ge and Huang 2011; Huang and Wang 2009). The IT service provider not only needs to decide short-term pricing but also needs to make the long-term capacity planning decision. We model this decision problem by the parameter “\( c \)” (the number of servers) in M/M/c queueing model. In other words,
IT service provider also maximizes objective function by choosing $c$. For each value of $c$, denote $p^*(c)$ as the optimal usage pricing for $c$ servers. The set of long-term profit-maximizing solution is denoted by $C^*$ and each optimal solution is denoted as $c^*$ because it is possible that there are multiple solutions in general. As a result,

$$ C^* = \arg \max_c \Pi(p^*(c), c) \quad (9) $$

**Preliminary Analysis and Results**

*Short-term Pricing under Per-unit Penalty Scheme*

As $c$ is the long-term decision variable, we omit $c$ in the analysis of short-term pricing problem for ease of exposition. To depict the parametric curve of $R(p)$ and $S(p)$, we need to first characterize $R(p)$ and $S(p)$. First, we already know that $R(p)$ is a concave, unimodal function from the literature. Second, we need to characterize $S(p)$ under the assumptions of eq.(3)-(5). It is straightforward to verify that when $S(p)$ is specified in eq.(3), $S(p)$ is a strictly decreasing function in $p$. With these two results, we can draw a parametric curve of $R(p)$ and $S(p)$ as shown in Figure 1. The vertical axis represents $R(p)$; while the horizontal axis represents $S(p)$. Each point on the parametric curve represents one feasible value of the $R(p)$ and $S(p)$ pair. The service provider can manipulate the location of the point on the parametric curve by adjusting the usage price $p$. Since $R(p)$ is first increasing and next decreasing in $p$ whereas $S(p)$ is decreasing in $p$, the direction of movement of $p$ is illustrated by the arrow in Figure 1. In other words, when $p$ increases, the point $(R(p), S(p))$ will move along the curve from the right to the left.

The service provider will adjust price $p$ to maximize its objective function specified in eq.(2). Given the parametric curve, our analysis can proceed with similar steps in microeconomics analysis about consumer choices or firm decisions. First, notice that the IT service provider prefers higher $R(p)$ and lower $S(p)$; point located closer to the upper-left corner is better. The parametric curve here is similar to the feasible set of consumer choices or firm decisions. By our objective function in eq.(2), we can draw a straight line, such as Line $L$ in Figure 1, that represents the iso-profit curve. Line $L$ is similar to the indifference curve for consumer choices. The optimal solution that maximizes eq.(2) is the intersection point between Line $L$ and the parametric curve. As a result, the optimal solution in Figure 1 is illustrated by $p^*$ whereas the benchmarking revenue-maximizing price (without considering the effect of $S(p)$) is illustrated by $p^{**}$.

We would like to highlight that the main contribution of our approach is: in standard microeconomics, the feasible set of consumer choices or firm decisions are exogenously given whereas in our parametric curve analysis, the feasible set of decisions ($p$ in this section) is derived by analyzing the properties of $R(p)$ and $S(p)$, which jointly determines the shape of the parametric curve. It is the “endogenous” parametric curve that characterizes the set of feasible solutions on the parametric curve. To the best of our knowledge, this analytical technique has not been proposed in the literature.

This geometric representation also provides an intuitive way to conceptualize the trade-off between $R(p)$ and $S(p)$, and this practical approach could be used as a decision making tools for managers who have no background in optimization: they only need to draw the parametric curve based on real numbers of $R(p)$ and $S(p)$. Next, the easiest way to find a solution is to find the intersection between the parametric curve and the 45 degree line that represents the objective function $R(p)$-$S(p)$ if $\theta=1$.

Mathematically, we can show the following lemma for characterizing the shape of parametric curve. Define $(R_{\min}, S_{\min})$ as the value of the point that is located closest to the origin.

**Lemma 1** If $[R_{\min}-R]/[S_{\min}-S]$ is decreasing (or increasing) in $p$, then the parametric curve is moving clockwise (or counterclockwise). As a result, $p^* < p^{**}$ (or $p^* > p^{**}$).
When $S(p)$ is specified in eq.(3), we can show that $\left[ R(p)-R_{\text{max}} \right]/\left[ S(p)-S_{\text{min}} \right]$ is a strictly increasing function of $p$. As a consequence, the direction of the parametric curve is always counterclockwise.

**Proposition 1** When the SLA penalty scheme is fixed per-unit, the optimal price with SLA is higher than the revenue-maximizing level: $p^* > p^{**}$.

**Corollary 1** (i) When the SLA penalty is fixed per-unit, a larger $\theta$ will lead to a larger optimal price $p$. (ii) The increase in per-unit penalty ($l$) will lead to increase in the optimal price $p$. (iii) if $((\partial^2\text{Pr}(w>k))/\partial k\partial p)>0$, the increase in the QoS threshold level ($k$) will lead to decrease in the optimal price.

Proposition 1 reveals the fact that when a service provider introduces SLA, it should increase the usage price to alleviate the congestion problem. Corollary 1 has important managerial implications for pricing and SLA contract design. First, it shows that the IT service provider needs to align these two decision making problems. Specifically, the more loss due to SLA (via $\theta$ or $l$), the higher the optimal price should be. Second, an increase in QoS threshold level $k$ should be accompanied with lower usage pricing.

Results for the SLA scheme in eq.(4) are qualitatively similar and are omitted to avoid repetition. The details of proof are also omitted due to page limit.

**Short-term Pricing under Average Penalty Scheme**

When $S(p)$ is specified as in eq.(5), there are two cases of parametric curves as shown in Figure 2. In either case, there are two straight lines: one line represents the scenarios that the service provider does not need to pay SLA penalty ($S(p)=0$) whereas the other line represents the scenario that $S(p)>0$.

In Figure 2(A), when the price is very small at point O, the average waiting time is for sure to be larger than $k$ in this asymptotic analysis. This scenario is illustrated by $\overline{OD}$ in Figure 2. As the price increases, both revenue $R(p)$ and SLA penalty $S(p)$ increase, and the parametric curve moves from O to A along $\overline{OD}$. As price increases, the average waiting time decreases because the demand is smaller. At point A, if the average waiting time becomes smaller than the threshold $k$, then $S(p)=0$. When $S(p)=0$, clearly, the parametric curve will become a straight line on the Y-axis, which is illustrated by $\overline{OC}$ in Figure 2. From point A, the parametric curve has a discontinuous jump to point B because that is the threshold value when the average waiting time equals $k$. At point B, when the price keeps increasing, then the revenue function will be first increasing and second decreasing. The parametric curve moves from B to C and next C to O.

In Figure 2(B), the discontinuous jump point from A to B occurs later. Similar to the case in Figure 2(A), the parametric curve starts from O to D and then from D to A because even when $R(p)$ starts decreasing, the average waiting time is still larger than the threshold $k$. As the price further increases, finally the average waiting time becomes smaller than $k$. At this point, the parametric curve jumps from A to B and next moves from B to O as price increases.
When we try to identify the optimal pricing by the iso-profit line, we can observe that the optimal pricing in Figure 2(A) is at point C. In Figure 2B, it is more difficult to determine the optimal solution. When the iso-profit curve is steep, the interception is at point B. When the iso-profit curve is flatter, the interception is at point C. Define price \( p' \) as the pricing point under which the average waiting time equals to the QoS threshold \( k \).

**Proposition 2** When the SLA includes an average penalty scheme specified in eq.(5), the optimal pricing depends on the values of \( p' \) and \( p^{**} \).

1. If \( p' \leq p^{**} \), IT service provider will choose optimal price the same as the expected-revenue-maximizing level. i.e. \( p^* = p^{**} \) and \( S(p^*) = 0 \). (Figure 2(A))

2. If \( p' > p^{**} \), there are two scenarios depending on the profit at \( p' \) (B in Figure 2(B)) and \( p^{**} \) (D in Figure 2(B)).
   a. If \( R(p') \geq R(p^{**}) - \theta S(p^{**}) \), the optimal price is \( p^* = p' \) and \( S(p^*) = 0 \).
   b. If \( R(p') < R(p^{**}) - \theta S(p^{**}) \), the optimal price is \( p^* = p^{**} \) and \( S(p^{**}) \) is the largest.

Note that although Case 1 and Case 2(b) of Proposition 2 result in the same optimal pricing (\( p^* = p^{**} \)), the SLA penalty are at two extremes. Specifically, if \( p' \) is smaller than \( p^{**} \), it implies that at the revenue-maximizing price, IT service provider can still meet the QoS in SLA without penalty. Clearly, it is also total profit-maximizing. In contrast, if \( p' \) is larger than \( p^{**} \), the IT service provider needs to consider an alternative pricing plan: should we increase pricing to relax congestion so to avoid paying hefty fine under SLA? This is the intuition behind comparing profit at D and profit at B in Figure 2(B).

The comparative statics in this case provide insights in several ways. First, it is straightforward to see that an increase in \( \theta \) in the objective function will only decrease the value of \( R(p^{**}) - \theta S(p^{**}) \), which is the RHS of the revenue condition in Case 2. Similarly, an increase in \( l \) (the penalty in SLA) will incentivize the service provider to avoid congestion by charging higher prices. Lastly, the QoS threshold \( k \) only has the effect on \( p' \); an increase in \( k \) means a less demanding QoS in SLA and will decrease \( p' \), which does not affect the solution in Figure 2(A). However, the value of \( k \) affects the value of \( p' \) and whether the solution is illustrated in Figure 2(A) or 2(B).

**Long-term Optimal Capacity Decision**

We extend the short-term pricing under the per-unit SLA scheme in eq.(3) to investigate long-term capacity problem. We slightly modify the revenue function by adding the fixed cost for IT infrastructure into \( R(p) \), i.e. \( R(p,c) = D(p) \times p - F(\mu,c) \). This section uses a simple example to illustrate our key findings in Figure 3.

In Figure 3, we only consider an example with \( c = 1, 2, \) or \( 3 \). Since we can derive the shape in the short-term analysis given any value of \( c \) (illustrated in Figure 1), conceptually, feasible solutions in this case consist of 3 parametric curves illustrated in Figure 3. When there are only 3 such curves, there are only
three nontrivial cases: 3 curves are “discontinuously” unimodal, convex, and concave, as illustrated in Figures 3a, 3b, and 3c, respectively. In all three cases, the curve on the right is depicted with $c=1$ whereas the one on the left is depicted with $c=3$. The reason is obvious: with a smaller $c$, the waiting time is longer and $S(p,c)$ is larger.

The optimal solution critically depends on the slope (i.e. $\theta$) of the iso-profit curve $L$ and the relative position of three parametric curves, which partly depends on the cost structure $F(\mu, c)$. As shown in Figure 3, the optimal solution must be on the convex hull ABC or ABCD of the 3 parametric curves. Figure 3A illustrates the first observation when the overall profit function is “discontinuously unimodal”. In this case, we can exclude all feasible solutions with capacity level that is smaller than the capacity that can generate the highest overall $R(p)$. For example, when $c=1$, all feasible pricing cannot become a potential candidate of optimal solutions. Figure 3B illustrates that it is possible that some intermediate values of capacity won’t be optimal when the overall profit function is “discontinuously convex”. In this case, the service provider either uses the lowest capacity or the largest capacity. Figure 3C illustrates the most common case: the overall profit function is discontinuously concave. In this case, all capacity levels could be optimal, depending on the value of $\theta$ of the iso-profit curve. All of these findings can be visualized and solved numerically by our parametric curve approach.

![Figure 3. Parametric Curve for Long-term Capacity Decision](image)

**Conclusion and Future Work**

This study proposes an innovative parametric curve approach to visualize and analyze the short-term pricing and long-term capacity problem for IT services under SLA. Figure 1 and 2 demonstrate how the penalty function in SLA may affect the optimal pricing decision. This research-in-progress also provides examples to illustrate how the optimal capacity planning could be affected by the infrastructure cost structure and service provider’s preference toward avoiding SLA penalty.

We plan to extend our analysis along several directions. First, we could include more decision variables (e.g. service rate $\mu$) and investigate the trade-off among pricing, capacity planning, and the service rate. Second, we can explore variations of the current model setup. For example, we could model the impacts of SLA parameters on the service demand directly. Third, we can examine more general pricing strategies. In this paper, we only examine linear pricing. We can extend the single parametric to multiple parametric by considering two-part tariff pricing or even the prevalent fixed-up-to pricing (i.e., the pricing plan used in the cell phone plans). With this generalization, we could investigate how demand uncertainty may affect different components in two-part or fixed-up-to pricing plans. Last, our geometric parametric curve approach could be applicable to other problems in which one decision variable affects two unimodal functions and the objective function consists of these two unimodal functions.

In summary, we believe this research could provide valuable insights and significant theoretical contribution to the understanding of the pricing, capacity planning and SLA contract of IT services.
References


