Pricing Strategy for Cloud Computing Services

Jianhui Huang, School of Information Systems, Singapore Management University, Singapore, jhhuang.2009@phdis.smu.edu.sg

Abstract

The cloud services market exhibits unique characteristics such as instant accessibility, fluctuating demand and supply, and interruptible service provision. Various pricing mechanisms exist in current industry practice, however, none is comprehensive enough to capture all these features. In my work, I identify key factors related to cloud computing pricing. My dissertation includes three essays. They employ multiple approaches, including market survey, game theory modelling, simulation, lab experiments and econometric modelling, to analyse the pricing strategy of cloud services vendors. The first essay highlights nine important factors in current cloud pricing practice and proposes three missing factors based on a market survey. In the second essay, I build an analytical model and use simulation to derive optimal pricing strategies for a monopoly cloud services vendor that operates in the reserved services market and the spot services market. In the last piece of work, I examine the client’s willingness-to-pay for customized cloud services through behavioural experiments.

Keywords: cloud computing, experimental methods, game theory, pricing strategy, simulation
1 INTRODUCTION

Cloud computing provides highly scalable Information Technology (IT) services to clients with instant access via the Internet. According to Gartner (2010), the revenue for the cloud services market was US$68.3 billion in 2010, representing a 16.6% increase from US$58.6 billion in 2009, and will reach US$148.8 billion by 2014. There are three main types of cloud services: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS).

Amazon is an IaaS vendor. It first introduced the Elastic Computing Cloud (EC2) in 2006 with usage pricing per instance-hour. Amazon later announced the EC2 reserved services and EC2 spot services in 2009. The EC2 reserved service requires client’s money commitment in advance, while the EC2 spot service implements an auction mechanism such that potential clients must bid for their desired resources and the price of resources changes over time (Amazon 2012). On the other hand, PaaS and SaaS vendors typically offer a package of IT service components and choose pricing tariffs similar to the charging plans of mobile phones. In practice, there is a trend toward turning computing services into small resource units, including CPU cycles, storage space, and I/O resource (Ben-Yehuda et al. 2012). More and more components of a computing instance can be purchased as a service individually. For example, data storage and backup, which were part of packaged services, now can be delivered as separate services. Targeted mainly to corporate clients, data services typically carry a monthly or yearly subscription fee.

Cloud services vendors are adopting a variety of pricing mechanisms, including usage-based fixed pricing, usage-based dynamic pricing, subscription-based pricing, reserved services contracts with a combination of usage-based fixed pricing and up-front fees, and auction-based pricing. In this study, I analyse different types of pricing strategies for cloud services. The well recognised literature on pricing information goods provides a good starting point (Maskin and Riley 1984; Varian 1995; Sundarajan 2004; Masuda and Whang 2006; Png and Wang 2010). I apply different research methodologies, including a market survey, game theory modelling, simulation, experimental design, and econometric modelling to answer the following research questions:

1. What are the major types of pricing methods currently used by cloud services vendors?
2. What are the key factors that should be considered in cloud pricing?
3. When and why should a cloud services vendor be interested in multiple pricing channels? To be more specific, how should Amazon’s EC2 mixed pricing strategy (with both fixed reserve pricing and dynamic spot pricing) be evaluated?
4. What are the key variables that will affect clients’ valuation of cloud services? How will they affect clients’ willingness-to-pay for customized cloud services?

My thesis includes three essays. The first is a market survey study for different cloud pricing cases. It identifies factors in pricing mechanism design, and makes recommendations to cloud services vendors to improve their pricing efficiency. The second study proposes a game theory model to study the profitability of offering multiple pricing channels by considering client self-selection behaviour. In the third essay, I conduct a two-factor behavioural experiment on a specially implemented online platform to examine variables that affect client willingness-to-pay for customized cloud services. The details of each work are explained as below.

2 PRICING PRACTICES IN THE CLOUD MARKET (STUDY 1)

The purpose of this study is to ground our understanding of existing pricing mechanisms that are currently adopted by the major cloud services vendors. I address four questions. First, what are major types of pricing methods currently used by major cloud services vendors? Second, what are the common pricing factors in the cloud services market? Third, what pricing factors affect service vendor’s pricing decisions? Fourth, what factors affect client adoption decisions? This study resulted in a pricing model that identifies common pricing components adopted in the cloud services market.
The study offers observations and recommendations on price-related factors that inhibit cloud services adoption.

2.1 Background Literature

Pricing practices in IT markets have shifted over the last two decades. Traditionally, CPU cycles, data storage and software applications have been sold as products, not services. Clients own the facilities after a one-time payment but have to pay for maintenance and upgrades. Most cloud services vendors have opted for “pay-as-you-go” pricing schemes. Clients are no longer bound to ownership and their payments are closely linked to real usage. The research shows that the existing pricing practice with complicated tariffs slow client adoption. Realizing that the accounting is very complex and the cost is unpredictable due to uncertainty about their usage patterns, customers slow down adoption of cloud computing services (Perry 2010). In order to tackle the complexity of the pricing of cloud services and its negative effects on adoption, a deeper understanding of the pricing mechanisms in the cloud services market is needed.

There has been a lot of discussion about the cloud business model (Buyya et al. 2009; Marston et al. 2011). Weinhardt et al. (2009) suggested that the pricing complexity of cloud services was one of the key challenges faced by vendors. Durkee (2010) investigated factors that affect pricing based on features characterizing cloud services, and argued that today’s price-focused cloud services market presents challenges for clients to choose appropriate services. Marston et al. (2011) reviewed several key cloud services vendors and identified key opportunities and challenges for the cloud industry. The recommendations for IS researchers include the following: First, optimal pricing strategy needs to be further studied in conjunction with capacity investment decisions and QoS guarantees; and Second, optimal pricing strategies should be examined for different cloud services market structures. Although the previous works have correctly recognized both the complexity and importance of pricing strategy for cloud services vendors, none has conducted careful investigations of cloud pricing.

2.2 The Analysis: Methodology and Findings

I looked at a total of ten cloud services vendors and seventeen different services provided by them. The seventeen cloud services contain seven IaaS, four PaaS, five SaaS offerings, and three brokerage offerings. Nine out of the ten vendors were from North America, and the other was from Switzerland. Pricing information was collected by visiting the vendors’ official websites. I followed the survey up with a review and a theoretical assessment of the literature related to the pricing of information goods.

My review identified nine pricing factors that are currently adopted by cloud services vendors. They are grouped into four categories: usage, reservation, technical support and penalty, as shown below in Table 1. I find that cloud services vendors tend to provide clients with flexibility for service type and technical support type, while they are not flexible for others. Definitions of each factor and their grouping information are given in Table 1.

<table>
<thead>
<tr>
<th>Four Factor Groups</th>
<th>Nine Pricing Factors</th>
<th>Definition</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Usage</td>
<td>Service Instance Type</td>
<td>Specifications of services, e.g., OS, size, location</td>
<td>Categorical</td>
</tr>
<tr>
<td></td>
<td>Unit Rate</td>
<td>Unit price of usage</td>
<td>$/unit</td>
</tr>
<tr>
<td></td>
<td>Total Usage</td>
<td>Total usage</td>
<td>Units</td>
</tr>
<tr>
<td>II. Reservation</td>
<td>Reservation Period</td>
<td>The length of reservation period (inclusive of Total Usage)</td>
<td>Hours</td>
</tr>
<tr>
<td></td>
<td>Reservation Fee</td>
<td>One-time payment for reserved service access</td>
<td>$</td>
</tr>
<tr>
<td>III. Technical support</td>
<td>Support Type</td>
<td>Characteristics of technical supports</td>
<td>Categorical</td>
</tr>
<tr>
<td></td>
<td>Support Charge</td>
<td>Periodic payment for technical support</td>
<td>$</td>
</tr>
<tr>
<td>IV. Penalty</td>
<td>Total Outage</td>
<td>The length of service down time</td>
<td>Hours</td>
</tr>
<tr>
<td></td>
<td>Compensation</td>
<td>The monetary penalty for vendor not fulfilling promises</td>
<td>$</td>
</tr>
</tbody>
</table>

Table 1. Definitions of the Nine Pricing Factors
A central observation from this review is that the current cloud pricing practice tends to be vendor-centric. Factors linked to service vendor’s interests are flexible, such as usage and reservations that directly affect the vendor’s benefits in terms of revenue and cash flow. On the other hand, the important factors linked to client’s benefits are fixed, such as the penalty terms that are crucial to clients but don’t contribute to revenue directly.

Based on the review, I derived three key recommendations. First, quantity discounts are missing in current cloud pricing practice and should be considered as a pricing factor for all categories of cloud services. Quantity discounts have been suggested as one of the most effective ways for vendors to segment clients, gain market power and obtain higher profits (Goldman et al. 1984; Monahan 1984). They also have been shown to be optimal for IT service vendors when there is no transaction cost (Maskin and Riley 1984). Second, the initial setup charge is missing and should be indicated in pricing. Third, penalty settings or service guarantee terms that guarantee promised quality level should be flexible enough and made negotiable to satisfy different types of clients. As shown in the market survey, a uniform type of service quality guarantee, including uptime and penalty terms, is set for all offerings from the same vendor. Such a practice ignores the natural differences between services, and is not appropriate. For example, mission-critical enterprise applications are costly for service downtime (Hiles 2005). So it is necessary to make those settings configurable.

3 MODELING RESERVED AND SPOT CLOUD SERVICES (STUDY 2)

Amazon first introduced usage-based fixed pricing for its cloud computing services in 2006, and later on, it announced a dynamic spot pricing mechanism in 2009. The spot services are novel because supply, demand and price for the spot resource change over time. A potential client must set a reservation price to compete for limited spot resources. He gets the service only if his reservation price is above the market price. In the case that the market price rises above his reservation price, this client’s service will be terminated by Amazon (2012).

There are strong needs from the client and vendor sides to understand this new pricing model. Clients need to figure out how to benefit from spot services, and make decisions on where to submit their jobs. Submission to the reserved services guarantees job completion but requires payment in advance. The spot services will be cheaper but there will be uncertainty related to job completion. The vendor also will need to understand the new pricing approaches to know whether it is wise to give clients more choices. My work proposes a game theory model to study this problem.

I ask three research questions. First, how should a cloud services vendor optimize reserved services pricing, in scenarios where spot services may or may not be offered? Second, how does the spot services market interact with the reserved services market? Third, how will the offering of spot services affect the vendor’s profit, client surplus and social welfare? This research is first to study interruptible services in a business context. I analyse the use of mixed pricing strategy for cloud services vendors, and offer insights for resource and capacity management for cloud services vendors.

3.1 Related Literature

Fixed fee pricing and usage-based pricing have been studied extensively. An interesting question is which pricing method is better. Maskin and Riley (1984) concluded that usage-based pricing is optimal for information goods with negligible marginal production costs. Sundararajan (2004) suggested that fixed-fee pricing, together with usage-based pricing, always outperforms pure usage-based pricing. Some other studies (Fishburn 2000; Sridhar et al. 2009) have found that fixed-fee pricing can outperform usage-based pricing in a competitive setting where the transaction cost plays an important role.

---

1 Another example of spot services is CloudSigma, which implements a special pricing mechanism using dynamically changing unit fee and bills every 5 minutes. The service is not interruptible once being launched (CloudSigma 2012).
Although spot services are of the same quality as other interruption-free services provided by Amazon, clients may perceive decreased quality due to possible service interruptions. Bhargava and Choudhary (2008) stated that it is optimal for a software vendor to offer a lower quality product when the variable production cost is low. Etzion et al. (2006) also showed that the optimal design of interacting sales channels, with auction and fixed-price selling, can outperform pure selling in most circumstances. This is due to disutility caused by waiting in the auction channel, that greatly lowers inter-channel competition, but meanwhile satisfies low value clients who originally would not make a purchase. A critical assumption is that high value clients perceive higher disutility induced by waiting in the auction channel compared to low value clients.

3.2 Dual Pricing Model for Cloud Services

I examined a cloud service market with fixed and dynamic pricing channels. The services are available to clients in two ways: they can purchase a fixed reserve price contract or pay a spot price which may change over time. This kind of a mixed setting has not been examined before. My game theoretical model examines the interplay between the fixed-price reserved services market and dynamic priced spot services market. I analysed the vendor’s pricing strategy relative to clients’ strategic job submission behaviour. I also evaluated the impact of the frequency of spot price changes. I will also analyse profit, consumer surplus and social welfare, and make suggestions to the cloud services vendor about how to do the best it can.

3.2.1 Model Set-up

A monopoly cloud services vendor offers its computing resources in two ways: as a reserved service and as a spot service. To use the reserved service, clients must buy a fixed price contract \((T, N)\) upfront. Clients pay \(T\) in advance to reserve \(N\) units of resources for future use. In such contracts, \(T\) is the reserved service price and \(N\) is the resource capacity limit. To use the spot service, the client does not need to pay up front. When the client’s job arrives though, he can only go to the spot market and pay the current spot market price, which will be changing as time passes. A summary of the model variables is given in Table 2.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T)</td>
<td>Fixed price for reserved contract</td>
</tr>
<tr>
<td>(N)</td>
<td>Resource capacity limit of reserved contract</td>
</tr>
<tr>
<td>(v)</td>
<td>Value of a single job</td>
</tr>
<tr>
<td>(v_l)</td>
<td>Lower bound of job value</td>
</tr>
<tr>
<td>(v_H)</td>
<td>Upper bound of job value</td>
</tr>
<tr>
<td>(\theta_l)</td>
<td>Probability of low spot price</td>
</tr>
<tr>
<td>(\theta_H)</td>
<td>Probability of high spot price, (\theta_H = 1 - \theta_L)</td>
</tr>
<tr>
<td>(p_L)</td>
<td>Low spot price</td>
</tr>
<tr>
<td>(p_H)</td>
<td>High spot price</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>Coefficient for clients’ sensitivity to service interruption</td>
</tr>
<tr>
<td>(\lambda_i)</td>
<td>Job arrival rate of client (i)</td>
</tr>
<tr>
<td>(\lambda)</td>
<td>Maximum job arrival rate of a client</td>
</tr>
</tbody>
</table>

Table 2. List of Model Variables and Parameters

In Amazon’s EC2 spot market, the spot service price varies, and price changes seem to be a random process (Ben-Yehuda et al. 2011). To capture this unique feature, I model the spot price as a random variable. By investigating historical spot price information, which can be accessed through Amazon Web Service Console (aws.amazon.com/console), I found that spot prices fluctuate often between a low base price and a relatively high price. Hence, I assume that the spot price will take on two values. One is a low price \(p_L\) with a probability of \(\theta_L\) and the other is a high price \(p_H\) with a probability of \(\theta_H = 1 - \theta_L\). I consider a \(k\)-stage game. The spot price will be dynamically adjusted at beginning of each stage. To simplify the analysis, I assume a job running as a spot service is subject to only one price change during its service duration. If the spot price increases compared to the last stage, the running
job will be terminated by the vendor; if the spot price stays unchanged or decreases compared to that in last stage, the job will not be affected and will execute to completion.

Each client faces uncertain demand, in terms of the number of jobs that need to use the cloud service. Client i’s job arrival follows a Poisson distribution with arrival rate λᵢ. Clients have heterogeneous demand. So some clients have a larger number of expected job arrivals than the others. λᵢ is located in a range of [0, λ̅] to represent this. The value of each job follows a uniform distribution over (vᵢl, vᵢh). When a running spot service is interrupted by the vendor, the client suffers and incurs disutility. The disutility is proportional to the job value v, denoted by −γv, with 0 ≤ γ ≤ 1. The coefficient γ measures a client’s sensitivity to service interruption.

3.2.2 Analysis

I consider a market with only one service as a benchmark case. So Benchmark Case I is that only reserved service is offered and Benchmark Case II is that only spot service is offered. In Benchmark Case I, I derive the optimal fixed price \( T^* \) for the reserved contract and show that the market is half covered.

In Benchmark Case II, I show that the vendor can control the parameter \( θ_L \) to gain higher profit. Note that \( θ_L \) measures the frequency of low spot price to appear. The optimal choice \( θ_L \) depends on two factors, the price ratio \( p_L/p_s \) and clients’ sensitivity to service interruption \( γ \). In general, \( θ_L \) should be increasing in \( γ \) when \( p_L/p_s \) is large and be set close to 1 otherwise.

I next analysing the dual market, in which both reserved service and spot services are available to potential clients. I derived the optimal reserved services price and compared it with the results from Benchmark Case I. The results show that the vendor should reduce its reserved services price when spot services are available, indicating the inter-channel marginalization effect. In addition, the vendor’s profit increases with \( γ \). This finding suggests that in a market with high sensitivity to service interruption, adding spot services will be likely beneficial to the vendor. Furthermore, a cloud services vendor can set \( θ_L \), the frequency of low spot price to appear, to optimize his profit. Using numerical examples and simulations, it is possible to demonstrate that the optimal \( θ_L \) depends on and increases in \( γ \).

Although the literature shows that multiple selling channels usually benefit the vendor, my analysis of the cloud services market suggests that it depends on the concrete market condition. In Study 2, I identify the situations under which offering dual pricing channels may outperform a single pricing channel for cloud computing resources. I further examine the market impacts of some important parameters, such as the client’s sensitivity level to service interruption, and the vendor’s spot price adjustment frequency – as well as the relationship between them. I now am analysing the changes in social welfare and consumer surplus when dual pricing channels are available.

4 A PRICING EXPERIMENT FOR CLOUD SERVICES (STUDY 3)

Empowered by advanced resource management technologies, cloud services vendors have been able to offer clients customized services. This study serves as a starting point for examining the possibility of an alternative pricing strategy for the cloud services market that allows customization. In this study, I ask: What are the key variables, and how do they affect client’s willingness-to-pay for customized cloud services? Following the experimental economics approach, I designed a laboratory experiment. I included two key variables, availability of risk analysis support and clients’ individual difference in risk propensity, in the research model.

4.1 Related Literature

Past research has examined service vendor’s pricing strategies under demand and service quality uncertainty. Paleologo (2004) noted that the traditional cost-based pricing is not value maximizing, in view of the dynamics of IT services adoption, due to the high level of demand uncertainty. Bhargava and Sundaresan (2004) suggested that service vendors are able to absorb the risk of fluctuations in
demand through economies of scale in service production. Hence, performance-based pricing is preferred by clients because of the unobservability of IT service quality and the lack of end-to-end control (Bhargave and Sun 2008). The uncertainty of service quality makes it difficult for clients to select appropriate price-quality contracts for the desired level of service performance.

Clients tend to prefer predictable to uncertain outcomes. This kind of preference can be explained by a client’s aversion to loss and asymmetries in the consequences. Prospect theory and asymmetric loss functions in behavioural economics and decision science study this (Kahneman and Tversky 1979; Weber 1994). A client’s risk propensity, described by a probability weighting function, describes his behavioural reaction toward losses and gains (Kahneman and Tversky 1979). According to prospect theory, people will be risk-averse when they perceive themselves to be able to gain, and will be risk-seeking when they are subject to loss, depending on the different perceptions they have relative to some reference point (Kahneman and Tversky 1979). When facing many options and the lack of data for a normal decision, a client’s decision-making will be inaccurate and biased, based on his beliefs and experience (Tversky and Kahneman 1974).

4.2 Research Model

Cloud clients’ willingness-to-pay for customized services provisions involves evaluation of alternative services: on-demand and spot provisions. The literature suggests that the client’s risk propensity and perceived risk information play important roles in their decision-making. I focus on examining the impacts of the client’s risk propensity and perceived risk information of spot service on their willingness-to-pay for customized cloud services. I use the following model as a basis for this study:

\[
WTP_{ij} = \beta_1 Price_{Reserved} + \beta_2 Price_{Spot} + \beta_3 RiskScore_i + \beta_4 RiskInfo_i + \beta_5 TaskDuration_j + \varepsilon
\]

where \(WTP_{ij}\) is client \(i\)'s willingness-to-pay for customized service provisions for running job \(j\); \(RiskInfo_i\) is the design variable that represents whether risk analysis information is available to client \(i\); \(RiskScore_i\) is the individual level measurement of client \(i\)'s risk propensity; \(TaskDuration_j\) is another design variable characterizing job \(j\); and finally \(Price_{Spot}\) and \(Price_{Reserved}\) are control variables, which stand for the unit price of spot services and reserved services. The error term \(\varepsilon\) stands for all other effects that are not explicitly included as right-hand side variables. I will add control variables such as sex and experience with cloud services in extended models that I plan to explore in the future.

4.3 Prototype and Experimental Design

I created a research prototype in the form of a hypothetical cloud vendor called SmarterCloud (www.smarter-cloud.biz) for this experimental study. The site allows a hypothetical client to compare and purchase cloud services. The experimental design is shown in Figure 1.

<table>
<thead>
<tr>
<th>Risk Analysis Support</th>
<th>(3 hours, 100 instances)</th>
<th>(5 hours, 100 instances)</th>
<th>(10 hours, 100 instances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>20 Subjects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>20 Subjects</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Experiment Design

For each setting specified in Figure 1, I need twenty subjects. Each setting includes three tasks which subjects will offer to a hypothetical cloud services vendor, with a set of parameters including price, compensation, and resource requirements. Two scenarios are presented. In one scenario, risk analysis support is offered. The other scenario does not offer such support; only historical price information is included. I conducted two rounds of pilot studies. The first round involved seven cloud computing professionals and the second round involves eight IS Ph.D. students and research staff mem-
bers. Feedback from the pilot studies was incorporated to validate and improve the experimental settings and the website interfaces.

The main experiment involved the participation of 46 staff members in a research institute in Singapore. As a qualifying condition, all participants had to have experience in business analytics and in using cloud services. To induce economic behavior, participants were given a participation reward of S$20 (S$1 = US$0.80) plus a performance-based bonus ranging from S$10 to S$50. After the subjects came to the experiment venue, they logged into the experiment website and were randomly assigned to one of the two treatments. After the experiment sessions, they were asked to complete a questionnaire to provide additional feedback on their perceived satisfaction and the usefulness of the website. Another set of questionnaires was distributed one week after the experiment to measure subject risk propensity, using instruments adopted from Weber et al. (2002).

5 SUMMARY AND FUTURE RESEARCH

Currently, I am extending Study 1 on pricing practices in the cloud services market to include more recent pricing practices from the cloud service market. Major information sources are Gartner and CloudSleuth.² Besides the empirical survey of the market, I will still develop a theoretical framework for cloud pricing to match the different layers and components of cloud services with client preferences for service reliability, cost predictability, and other concerns.

Study 2 on the analysis of fixed reserved versus dynamic spot services is ongoing. Though I have built a baseline model and solved it, I still need to derive more general results. I will extend the model in multiple ways. First, I have not incorporated the effect of the capacity limit specified by the reserved services contract. What will happen if the vendor puts a capacity limit on clients? How will client contract purchases and job submission strategies change? Second, I need to analyse the vendor’s decision problem in an integrated way. A cloud services vendor’s decisions related to the reserved price, the resource capacity limit, and the spot price adjustment frequency are influencing one other. In my analysis, however, I have separated their effects. I will consider an integrated solution, which can be implemented through numerical analysis. Finally, I will provide more insightful discussion and interpretation.

Study 3 on pricing customized cloud service provisions is still in progress too. I plan to conduct an ANOVA test to examine the effect of risk information analysis support on client willingness-to-pay. I will also use an econometric model to assess the relationship between client risk scores and their willingness-to-pay. I will further investigate the interaction effects between the two variables and the weighting function for risk propensity with different job settings in the simulation.

References


² CloudSleuth (cloudsleuth.net) is a partner driven community closely follows the development of cloud computing industry and active in disseminating information contributed by partners.


