2-2014

Pricing Strategy for Cloud Computing Services

Jianhui HUANG
Singapore Management University, jhhuang.2009@phdis.smu.edu.sg

PRICING STRATEGY FOR CLOUD COMPUTING SERVICES

HUANG JIANHUI

SINGAPORE MANAGEMENT UNIVERSIRY

2013

PRICING STRATEGY FOR CLOUD COMPUTING SERVICES

by

HUANG JIANHUI

Submitted to School of Information Systems in partial fulfillment of the requirements

for the Degree of Doctor of Philosophy in Information Systems

Dissertation Committee:

Dan Ma (Supervisor/Chair)
Assistant Professor of Information Systems
Singapore Management University

Robert J. Kauffman
Professor of Information Systems
Singapore Management University

Lin Mei
Assistant Professor of Information Systems
Singapore Management University

Yang Yinping (External Reviewer)
Scientist and IHPC Independent Investigator
Institute of High Performance Computing (IHPC)
Agency for Science, Technology and Research (A*STAR)

Singapore Management University
2013

Pricing Strategy for Cloud Computing Services

HUANG JIANHUI

Abstract

The cloud computing 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. None of these pricing mechanisms, however, is comprehensive enough to capture all these features in a way that allows the vendors to optimize resource allocation. This dissertation identifies key factors related to cloud computing pricing, and examines their interplays.

This research uses multiple approaches, including a market survey, game theory modeling, simulation, lab experiments and econometric modeling, to analyze the pricing strategy of cloud services vendors. A field study of a representative set of cloud vendors' pricing approaches provides background information to motivate this dissertation. Important factors in current cloud pricing practice are highlighted. This part of the work offers an overview of how cloud computing services are priced and how the pricing approaches have evolved over time. This is important for identifying the causes of confusion among clients regarding the pricing of cloud computing services, especially when new pricing methods emerge.

The following study reports on an analytical model and simulations to derive optimal pricing strategies for a monopoly cloud services vendor that operates in the reserved and spot-price services market. Although interruptible spot services potentially cause undesired consequences for clients, the cloud services vendor can improve its profit by offering low quality interruptible spot-price services, together with high quality fixed-price reserved services.

The last study examines clients' willingness-to-pay for brokered cloud services offered with a hybrid pricing mechanism in the presence of fixed-price reserved and spot-price on-demand services through behavioral experiments. This study yields two findings. First, subjects' willingness-to-pay is affected by their informedness of the service interruption risk associated with spot-price on-demand services. Second, this effect is moderated by their aversion to risk.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am indebted to the members of my dissertation committee, Ma Dan, Rob Kauffman, Lin Mei, and Yang Yinping for their invaluable guidance and support on my study, research, and life. I sincerely appreciate their enormous help on the preparation of my thesis.

I also thank Richard Shang Di for many inspiring and constructive comments and suggestions. I especially thank the Institution of High Performance Computing (IHPC), Agency for Science, Technology, and Research (A*STAR) for funding the user study on “Cloud Computing Pricing Mechanisms”.

Dedicated to my parents

1 Introduction

There are the controversial viewpoints on what is cloud computing. Detailed discussions on this issue can be found in Vaquero et al. (2009) and Madhavaiah et al. (2012). Cloud computing provides highly-scalable IT services to clients with instant access via the Internet (Armbrust et al. 2010) while offering enterprise clients business agility and cost efficiency. Economies of scale, which drive down the cost of services provision and raise vendor profit, make cloud computing services financially attractive and the cloud business sustainable (Armbrust et al. 2010, Foster et al. 2010, Marston 2011, Bardhan et al. 2010). Cloud computing services are transforming enterprise IT provision and gaining popularity. According to Gartner (2012, 2013), revenue in the global cloud computing services market was US$111 billion in 2012, representing a 21.4% increase from US$91.4 billion in 2011, and it will reach US$206.6 billion in 2016.

Cloud computing services appeal to clients with an unknown or changing demand for computing power and large batch processing tasks (Armbrust et al. 2010). As prices are constantly being driven down (Stevens 2012, Heath 2013), in the long run, it is likely that clients will rely on cloud computing for all IT-related services. This is because it is more economical than in-house systems, especially for clients who mainly execute data-intensive computing tasks, where the cost savings can be up to 95% (Kondo et al. 2009). This will have a great impact on businesses as reducing cost and increasing profit are major concerns. We have already seen cloud computing services playing revolutionary roles in several industries. The cost reduction of cloud computing services has had a dramatic impact on the software industry that is similar to the

impact of semiconductor manufacturing on the hardware industry (Armbrust et al. 2009). It has created significant opportunities for IT startups with limited initial capital to invest (Etro 2009). Music, movies, and television series are all now offered via online streaming services, backed by cloud services, such as Netflix. In the healthcare industry, companies such as Practice Fusion, Microsoft, Qualcomm Life, Philips, Verizon and AT&T have launched cloud-based vertical solutions that support greater health data sharing and accessibility. This will greatly enhance the efficiency of collaboration between healthcare providers and ensure seamless, personalized healthcare services (Grindle et al. 2013).

The idea of obtaining computing power on demand is not totally new though. It was first explored by researchers in the grid computing literature of the mid-1990s. They studied how to make the idle computer resources available all over the world to those in need of computing power, with information technologies that coordinate clients to discover, request, and use computer resources in an on-demand manner (Foster et al. 2008).

Despite their similarities in vision, the grid and the cloud have crucial differences in their business models. There has been a surge of adoption of cloud computing services in the past several years, which grid computing was not able to achieve after more than 15 years of development. In the grid computing business model, lack of control of underlying infrastructures forced vendors to face the uncertainty of resource availability and over-provisioning (Foster et al. 2010). This posed the problem of the coordinator being unable to deliver the required levels of quality of service (QoS) (Broberg et al. 2008, Buyya et al. 2007). In the cloud computing business

model though, a cloud services vendor has full control of its infrastructure, which reduces the uncertainty of resource availability and over-provisioning of services. In this respect, the dedication of computers to cloud services is pivotal to the success of the cloud business model. It releases the vendors from the concern of obtaining sufficient computer resources, and enables them to focus on the design and pricing of the cloud services offerings in order to make sufficient profit.

Although cloud computing has dedicated servers that support services provision, maintaining QoS and meeting service level agreements (SLAs) are still challenging. Cloud computing services are delivered via the Internet, which lacks a QoS guarantee mechanism. In addition, potential system and software failures that may shutdown the cloud platform are more difficult to detect in large-scale distributed systems like the cloud (Armbrust et al. 2010). In addition, demand uncertainty is also a challenge for cloud services vendors (Henzinger et al. 2010). Furthermore, heterogeneous performance of virtual machines increases the difficulty of vendors in maintaining QoS and SLA (Mei et al. 2011, Jayasinghe et al. 2013, Ou et al. 2013, Pu et al. 2013). These challenges motivate researchers to study SLA-based resource-allocation mechanisms in cloud computing (Buyya et al. 2009, Benaroch et al. 2010, Wu et al. 2013, Zhao et al. 2013).

One such example is the work by Sim (2010). The author proposed a complex negotiation mechanism for the cloud to enable dynamic SLA negotiation between: (1) a cloud broker and its clients; and (2) a cloud broker and other cloud services vendors. The market mechanism potentially improves the overall market efficiency as clients get tailored services. The dimensions of the negotiable SLAs, however, are not dis-

cussed in a real business context. In addition, an understanding of how clients value flexibility in different aspects of cloud services provision—such as time of access, duration of use, computing capability required, and others—is important for the market mechanism to be viable. The second part of this dissertation pursues insights into client characteristics.

Besides grid computing, cloud computing also shares similarities and differences with several other products and services. Regarding the cost of providing services, cloud computing services have similar characteristics to information goods and traditional capacity-limited services, such as hotel room and airline ticket bookings. Information goods require a large investment in the creation of the first copy, while the production and distribution costs for each additional copy are very small. Analogously, the marginal cost of providing a single unit of cloud computing services is also small, if not negligible, compared to the huge investment in building the infrastructure.

The hotel and airline industries have to balance their operational capacity with fluctuations in demand. It is the same in the cloud computing services industry. On average, however, the frequency of repeated purchases of cloud services is far higher than that of hotel rooms or airline tickets: frequent travellers probably book hotel rooms and purchase airline tickets several times a month, while enterprise clients use cloud services every day. Furthermore, the duration of usage relative to the amount of billable time that is used is longer for cloud computing services than for hotel rooms, for example.

Nevertheless, given their similarities, it is worthwhile to ground our understand-

ing of cloud services in the literature related to information goods, the hotel industry, and the sale of airline tickets. Knowledge from research in those areas is useful. The lack of standardization of cloud services contracts and delivery requires clients to be knowledgeable though. Uncertain standards make it difficult for clients to compare different services and make purchase decisions. This also creates difficulties for interoperation among different cloud services.

In the market, initially there were three main types of cloud computing services: *infrastructure as a service* (IaaS), *platform as a service* (PaaS), and *software as a service* (SaaS). As the cloud computing services market matured, more categories of services have emerged, such as *data storage as a service* (DSaaS), *hardware as a service* (HaaS), *desktop as a service* (DaaS), *business process as a service* (BPaaS), and many more (Rimal et al. 2009). Cloud services vendors have adopted 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.

Among the various pricing mechanisms, usage-based pricing is one of the main selling points of cloud computing services. With usage-based pricing, clients pay the vendor on a consumption basis. This is like when utility companies charge clients for basic utilities such as electricity, water, and gas. Usage-based pricing has various forms. Usage-based fixed pricing is implemented by vendors, who charge by the hour, minute or second. Usage-based dynamic pricing is implemented differently. Some vendors let the price fluctuate while maintaining a level of service quality. Others may degrade service quality by restricting the service capability that a client can

choose, or by introducing an interruption mechanism through which the vendor can hold back resources without notifying its clients (Amazon 2013).

One good example is Amazon. As an IaaS vendor, Amazon first introduced the Elastic Computing Cloud (EC2) in 2006 with usage pricing per instance-hour. After that, Amazon announced its EC2 reserved services and EC2 spot services in 2009. The EC2 reserved services require the client's financial commitment in advance, while the EC2 spot services are implemented with an auction mechanism. This means that potential clients must bid for their desired resources and the price of the resources changes over time (Amazon 2013). The various pricing mechanisms Amazon implemented for its cloud computing services creates difficulties for clients in estimation, comparison, and optimization of the cost to use Amazon.com's cloud computing services. Prior research has investigated this problem and suggested to formulate the cloud services consumption plan as a stochastic programming problem to achieve cost minimization (Chaisiri et al. 2012), or to use demand prediction techniques such as Kalman filter to enable adaptive purchasing of cloud computing services and minimize cost (Hwang et al. 2013).

On the other hand, PaaS and SaaS vendors typically offer a package of IT services and choose prices similar to charging plans for mobile phones. Some PaaS vendors also implement pricing strategy similar to IaaS vendors. Their prices are different for different types of service instances, but the underlying mechanism is task-oriented.

With cloud computing services, clients can access virtualized hardware or operation systems and software applications of all kinds via cloud computing platforms, with capital expenses that are much lower than before. In the traditional business

model of IT services, selling and licensing are major means of revenue for the vendors. The pay-as-you-go business model thus puts considerable competitive pressure on traditional IT service vendors, including both hardware and software vendors such as IBM, Oracle, Hewlett-Packard and many others. Meanwhile, some traditional IT services vendors are also entering the cloud market. Some examples are Applications Cloud from Oracle, SmartCloud from IBM, and Windows Azure from Microsoft.

In practice, various implementations of pricing for cloud services have caused some uncertainty among clients, diminishing their interest in adopting cloud computing services. This also may lead them to misalign their business goals with the cloud services they have adopted. Such misalignment creates the possibility that core business functions are run on cloud computing services that are cost-effective to some extent but subject to unexpected service termination by vendors, such as Amazon EC2's spot-price services. The consequences can be critical (Howard 2011).

This dissertation seeks answers to why cloud vendors have adopted so many different pricing mechanisms, other than differentiate their services from those of other vendors (Choudhary 2010). What is the best strategy if a cloud vendor implements more than one type of pricing scheme?

The large price reduction associated with interruptible cloud services offers the client incentives to bear the risk of unexpected services interruption. Statistical models of spot prices are used to predict the payment for jobs running as spot instances (Javadi et al 2011). Technologies, such as checkpointing, can be employed to reduce the impact of service interruption to an acceptable level (Yi et al. 2010, Yi et al. 2013). These technologies, however, are yet to mature. Meanwhile, different parties

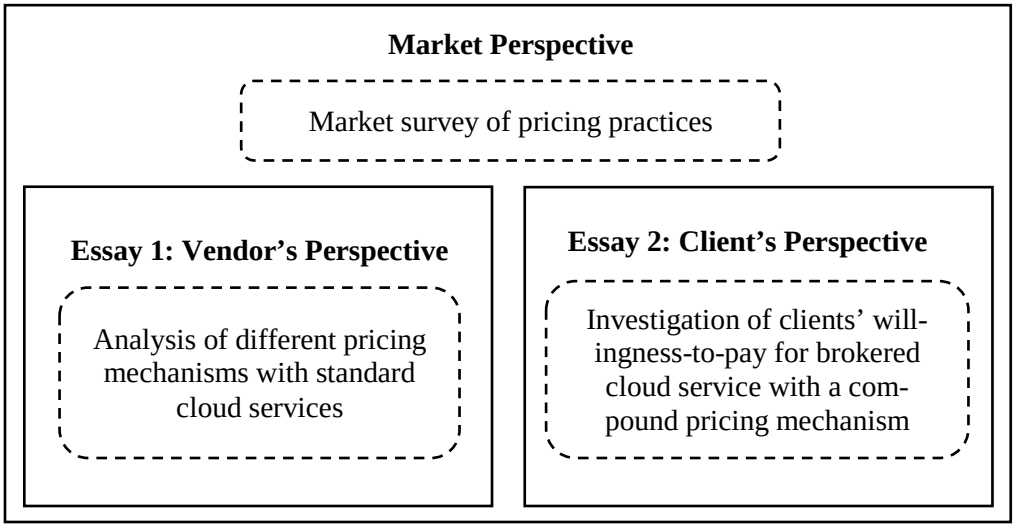
are entering the market to address this problem for clients. Emerging cloud brokerage services offer clients low-cost, immediate-restoration-upon-interruption cloud services. The services are suitable for highly-parallel computing tasks of limited duration, but are unsuitable for tasks that require high availability. So far there has been no discussion about the value of services interruption mitigation technologies, which is critical for their commercialization. Neither have issues related to the economic impact of cloud brokerage been explored scientifically.

Cloud services brokers usually build a software stack on top of their IaaS offerings. They also charge clients in a similar way to major IaaS vendors. Is this a good practice for them though? Why would a cloud broker operate on one layer but not on other layers in the cloud computing services ecosystem? How will brokerage services affect existing cloud services and their vendors?

Prices signal the quality of services. And a vendor's pricing scheme directly affects a client's decision on whether to use the services. For both cloud services vendors and brokers, identifying the common factors in cloud services pricing is critical, which is one of the key issues in this dissertation. By identifying the common factors in cloud services pricing, the findings will contribute to the improved implementation of cloud services. These common factors reflect client and vendor concerns about cloud services from different perspectives, including the financial perspectives of risk management and valuation.

The goal of this research is to enhance the understanding of cloud computing services pricing and its underlying economic mechanisms. Figure 1.1 gives an overview of the dissertation.

Figure 1.1 Research model



This dissertation will include three parts. The first is a market survey with cases involving different cloud pricing approaches. It provides an overview of pricing practices in the cloud computing services market, and identifies important factors in current cloud pricing mechanism designs. The second is developed with a cloud services vendor's perspective, proposes a stylized analytical model to study the profitability of offering cloud services with multiple pricing methods by considering clients' self-selection behavior. The third conducts a behavioral experiment on a specially-implemented online platform to examine variables that affect client willingness-to-pay for brokered cloud services. It includes guaranteed job completion, and a hybrid pricing mechanism.

The rich literature on pricing information goods (Maskin and Riley 1984, Varian 1995, Sundarajan 2004, Masuda and Whang 2006, Png and Wang 2010) and revenue management (Boyd and Bilegan 1999, McGill and Van Ryzin 1999) provides a good starting point. This dissertation employs a number of research methodologies, including a market survey, analytical modeling, simulation, experimental design, and econ-

ometric modeling to answer the following research questions:

(1) What are the key factors that should be considered in pricing cloud computing services?

(2) Should a cloud vendor be interested in multiple pricing approaches? Related to this, how should Amazon.com's EC2 hybrid pricing strategy, with both fixed-price reserve pricing and dynamic spot pricing, be evaluated?

(3) What are the key variables that will affect clients' valuation of cloud services? How will they affect clients' willingness-to-pay for brokered cloud services with guaranteed job completion?

The rest of the dissertation is organized as follows. Chapter 2 offers an overview of the pricing practices in the current cloud computing services market. Chapter 3 presents the model of a monopoly cloud computing services vendor offering its services with fixed pricing, spot pricing, or hybrid pricing that contains both. Chapter 4 presents an experimental study of client willingness-to-pay for a brokered cloud computing services with a job completion guarantee, in the presence of fixed pricing and spot pricing. Chapter 5 contains my reflections on the IRB permission process for this research. Chapter 6 concludes the dissertation and discusses limitations and future research.

2 Pricing Practices in the Cloud Computing Services Market

This chapter presents an overview of pricing practices in the cloud computing services market. Specifically, it discusses a wide range of cloud computing services and compared their pricing schemes to identify common factors that affect total price. I also describe several cases where a noticeable trend—of how cloud services vendors change their pricing decisions and services offerings—is observed. I also will summarize the observations and discuss their implication for both research and practice.

2.1 Background

Cloud computing services vendors are employing a number of pricing mechanisms, such as usage-based, subscription-based, and a hybrid mix of fixed and usage-based pricing. Even within a specific type of pricing though, variation exists in the market. For example, subscription plans for cloud services can differ in multiple dimensions, including the length of the subscription period (monthly, quarterly, or yearly), the number of clients as system users, the number of hosted applications, and so on. It is interesting to see such diversity in pricing approaches in the marketplace. And there are conflicting viewpoints, in terms of cost, performance, compliance, and management, about whether cloud computing services are a better alternative to in-house systems.

Despite the significant marketing hype, the wide-spread adoption of cloud infrastructure and services by organizations is yet to materialize. Comparison studies have been performed to evaluate the option of using the cloud versus its alternatives. This work has compared cloud services to desktop grids in scientific data-intensive applications (Kondo et al. 2009), cloud and community clusters in parallel MPI applica-

tions (Walker 2008), and computational and storage costs for a montage application deployed in the cloud (Deelman et al. 2008). Cost and performance comparisons are the key concerns of the studies in this line of research. They also highlight the lack of clarity about the cost and performance of cloud computing services. Prior research has pointed out that the monetary cost of running scientific data-intensive applications using Amazon.com's S3 data storage services is out of reach for some potential clients, because the storage services—including availability, durability, and access performance—can be expensive but not altogether necessary (Palankar et al. 2008). Contradictory results were presented by Deelman et al. (2008): with storage cost reductions, using cloud services is cheaper than in-house systems for data-intensive applications.

Many of the doubts are rooted in the fact that clients are unclear about the total cost of adopting cloud computing services (Durkee 2010). The complicated price structures are the key challenge for cloud services vendors (Weinhardt et al. 2009), which have slowed down the adoption of cloud computing services (Perry 2010). In addition, it is not easy for an individual client to monitor its cloud services usage, adding to the uncertainty in cost of using such services. This motivated me to conduct a comprehensive market survey on cloud computing services pricing mechanisms and to explore the underlying rationale for why they are offered.

It is appropriate to review the different types of pricing methods currently used by major cloud computing services vendors, and identify the common elements for these different pricing methods. Although previous studies that review cloud services vendors' offerings have recognized the complexity and importance of appropriate pricing

strategy (Durkee 2010, Marston et al. 2011, Demirkan et al. 2008), none of them conducted a careful investigation of pricing strategies for the different service layers. The overview of cloud service pricing complements these studies and provides a basis for further development of this dissertation. Furthermore, it helps to identify factors that reflect vendor concerns when they make pricing decisions and evaluate cloud services, but are missing in current pricing practices.

I examined 19 cloud services vendors and 27 services offerings they provide, including four major types of cloud computing services: IaaS, PaaS, SaaS, and brokered cloud services. IaaS delivers computer infrastructure based on virtualization technology. PaaS has an additional layer on which clients can run applications without knowing how its underlying infrastructure is implemented. SaaS provides application services functioning as locally-installed software (Vaquero 2008). Brokered cloud services are provided by cloud services brokers who operate as intermediaries, aggregators, or arbitrageurs (Gartner 2010).

2.2 Data

I collected pricing information related to the major players in the cloud computing services market. The criteria for the selection of a vendor include: (1) the vendor must make pricing information on all its services available on its official web site; and (2) the vendor must have been selected at least once for review in Gartner's Magic Quadrant Report (Leong and Chamberlin 2010, 2011; Leong et al. 2012, 2013). The series of reports lists cloud computing services vendors that are leaders in the market, in terms of revenue and market share. This is to ensure that the sample was a fair representation of the vendors in the market.

From this process, I obtained 19 cloud computing services vendors that offered 27 types of services, including 15 IaaS, 6 PaaS, 7 SaaS, and 3 cloud services brokerage services. (See Table 2.1.)

Table 2.1 Cloud services vendors selected for inclusion in this research

Service type	Service name	Vendor name	URL
IaaS	Amazon EC2 On-Demand Instances	Amazon	goo.gl/fEzlD
	Amazon EC2 Reserved Instances	Amazon	goo.gl/fEzlD
	Amazon EC2 Spot Instances	Amazon	goo.gl/fEzlD
	Amazon S3	Amazon	goo.gl/BcG1n
	Infrastructure-as-a-service	Alatum	goo.gl/w0B9d
	Enterprise VM Hosting	nGrid	goo.gl/ihEuI
	CloudSigma	CloudSigma	goo.gl/20mev
	Cloud Servers	GoGrid	goo.gl/6Z4bO
	Joyent Cloud	Joyent	goo.gl/xkcwA
	Rackspace Cloud Servers	RackSpace	goo.gl/cSZEA
	FlexiScale public cloud	FlexiScale	goo.gl/I9rwE
	IaaS	Profit Bricks	goo.gl/weH6L
	Google Compute Engine	Google	goo.gl/RehH4
	HP Cloud	HP	goo.gl/ZV3Fo
	CloudLayer Computing	SoftLayer	goo.gl/8VKj3
PaaS	Google App Engine	Google	goo.gl/RLtG8
	CloudFare	CloudFare	goo.gl/Jqt9Q
	Force.com	Salesforce	goo.gl/Lo8jj
	Microsoft Windows Azure	Microsoft	goo.gl/rDwP5
	Microsoft SQL Azure	Microsoft	goo.gl/rDwP5
	Amazon Beanstalk	Amazon	goo.gl/Tpu0E
SaaS	Service Cloud	Salesforce	goo.gl/7sjJf
	Sales Cloud	Salesforce	goo.gl/PkojZ
	Chatter	Salesforce	goo.gl/g7Lqq
	Jigsaw	Salesforce	bit.ly/g6i6Um
	Google App for Business	Google	goo.gl/kxkeZ
	NetSuite Financial Management	NetSuite	goo.gl/dtqTH
	Office 365	Microsoft	goo.gl/Au3tM
Cloud Brokerage	PiCloud Public Cloud	PiCloud	goo.gl/JGbKT
	RigtScale Cloud ComI Editions	RightScale	goo.gl/PDDwl
	Integration Cloud	Boomi	goo.gl/oO3lz

All information was collected from vendors' official websites in October 2013. Vendors may have changed their website structures and content related to its services

pricing over time.

2.3 Observations

Pricing Innovations

Since 2006, Amazon has been innovative in its cloud services offerings and pricing. It first introduced its Elastic Compute Cloud (EC2) services in 2006 and used an on-demand service provision model with a pay-per-use pricing mechanism. Payments were based on actual usage and charged in hours. After the services were announced, the per-hour price was not changed frequently. This is pay-per-use pricing or fixed-price on-demand pricing. A series of pricing innovations from Amazon and its competitors is presented here. (See Figure 2.1.)

Figure 2.1 Pricing innovations

In 2009, Amazon announced two other new services offerings: EC2 reserved instances and EC2 spot instances. For consistency, let us call these *fixed-price reserved services* and *spot-price on-demand services*. Fixed-price reserved services use fixed-price reserved pricing. A client must pay a fixed fee up front to reserve a unit of the services. The client still needs to pay for actual usage. But the per-hour rate was low-

er than that in the pay-per-use model Amazon introduced in 2006 though. Spot-price on-demand services use a different pricing model that is auction-based. The major difference between spot-price on-demand services and the other options Amazon offered was that the spot-price services were subject to potential interruption. This novel pricing mechanism allowed Amazon to ration its idle computer resources based on client willingness-to-pay for the services.

Amazon is innovative in services and pricing design, however, it has locked into a specific billing cycle: it always charges clients by the hour. In 2011, a Zurich-based IaaS vendor announced a burst-pricing scheme that had a billing cycle as short as five minutes. This is similar to the practices of telecommunication companies that initially offer monthly subscription plans only, and then start to offer per-second billing. It is likely that the cost associated with metering and billing in such a short interval has been driven down. In 2012, a cloud computing services broker, PiCloud, offered its clients even more value by providing a usage-consolidation service. A client can use 1,000 compute instances, each active for one second only, and then pay the price for using one compute instance for 1,000 seconds. The same usage would have cost the client 1,000 instance hours using Amazon EC2.

Pricing Factors

A typical cloud computing service offering is characterized by five aspects: *service type, validity period, technical support type, service quality guarantee,* and *penalty terms.* Detailed pricing information is summarized in Appendices A1 to A4. Cloud computing services of different types may have different price structures though. For example, an IaaS offer might have many service components that are

charged for separately, for example, Internet usage and storage. But a SaaS offering may be charged with a single price for a bundle of service components.

Based on my observations, vendors tend to give clients more choices in the type of service and technical support. Let us call them the “flexible parts” of cloud services offerings. For example, IaaS and PaaS vendors provide a rich set of computing services with varied capacity, operation systems, and locations (see Appendices A1 and A2). SaaS vendors and cloud brokers offer different service bundles (see Appendices A3 and A4). For each service type, there are multiple choices for technical support. For the other three aspects, however, vendors provide limited options. These are the “inflexible parts.” For example, whatever type of service is chosen by a customer, the penalty term is always identical. Amazon EC2 provides a 99.95% uptime guarantee but offers the same penalty term for all computing instances. So do Microsoft Windows Azure, Google App for Business, and Google Compute Engine.

2.4 Nine Factors That Describe Cloud Computing Services Pricing

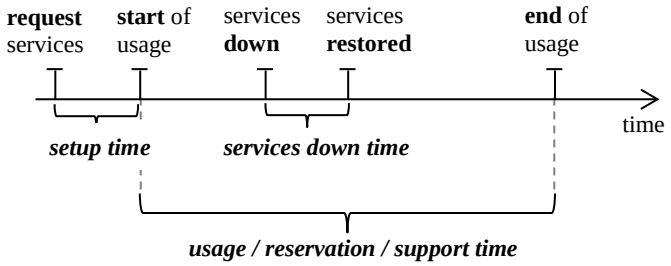
Based on the market survey, there are nine pricing factors common to the various types of Cloud services. They fall into four groups: usage, reservation, technical support and penalty. (See Table 2.2.)

Table 2.2 Definitions of the nine pricing factors

Factor Groups	Pricing Factors	Unit
I. Usage	Specifications of services, including OS, size, location	categorical
	Unit price of usage	$/unit
	Total usage	units
II. Reservation	The length of reservation period	Hours
	One-time payment for reserved privilege of using the service	$
III. Technical support	Characteristics of technical supports	categorical
	Periodic payment for technical support	$
IV. Penalty	The length of service down time	Hours
	The monetary penalty for vendor not fulfilling promises	$

In order to clearly discuss the pricing factors, it helps to understand the important time points in a complete cloud computing services transaction. (See Figure 2.2.) A client will first request a cloud services vendor for access to services. The vendor needs to process the request and setup the services for the client. After this *setup-time*, the client will be able to start using the cloud computing services. The *usage time* will count from this point in time until the client ends its use of the services. I call this period *reservation time*, because the resources are reserved for the client's use. Assuming the client will experience a period of services being down. The period when the time the services go down to when the services are restored is the *services down time*.

Figure 2.2 Cloud services transaction



Usage-based Pricing

Usage-based pricing is optimal for information goods with negligible marginal production costs (Maskin and Riley 1984). Most IaaS vendors examined in this study use only usage-based pricing. Some of them use a combination of fixed monthly subscription fee pricing and charges clients for overage with usage-based pricing though. The prevalence of usage-based pricing among IaaS offerings is inconsistent with the findings in Fishburn et al. (1997) and Sundarajan (2004), but it is consistent with the findings in an earlier study by Maskin and Riley (1984). The key difference in these studies is whether the transaction costs associated with usage-based pricing are negligible. IaaS vendors commonly implement a highly-automated management system, which generally has very low transaction costs. It is reasonable for IaaS vendors to adopt this kind of pure usage-based pricing scheme.

Prior research has suggested that fixed fee pricing, together with usage pricing, always outperforms pure usage-based pricing (Sundarajan 2004). Such two-part pricing is always no worse than any nonlinear pricing strategy (Masuda and Whang 2006; Png and Wang 2010). These findings are consistent with the pricing practices in the cloud market. Many PaaS and SaaS vendors have adopted the two-part tariff model. (See Appendices A2 and A3.) Clients pay a subscription fee for default usage quotas and also pay an additional price if the usage exceeds the pre-assigned limit.

Usage-based pricing are affected by the type of services and length of the usage period. Vendors usually offer fixed usage quotas for subscription plans. If a client uses more than its usage quota, usage that exceeds the quota will be charged based on the extra usage and unit price. The additional payment will be the product of the extra

usage and unit price.

Reservation-based Pricing

Most PaaS, SaaS, and cloud brokerage services vendors offer subscription plans (Appendices A2, A3, and A4). Some IaaS vendors provide similar plans too. For example, Amazon EC2 offers clients with reserved compute instances for a period of 1 year or 3 years, as shown in Appendix A1.

Reservations typically are associated with increased revenue in the restaurant and hotel industry (Alexandrov and Lariviere 2008). In the case of hotel reservations, rooms usually are scarce in popular attraction areas and travelers will be willing to pay for reservations. This rationale does not hold in cloud services though. Computing capacity is expandable at very low cost. So clients have little incentive to reserve services if there is no reward (Meinl et al. 2010). From the vendors' point of view, reservations benefit them by reducing their demand uncertainty because reservations lead to more predictable demand compared to random walk-ins. Any pre-paid reservation fees can potentially enhance a vendor's cash flow, and generate lock-in with clients. Therefore, vendors should incentivize clients for reservations by offering them favorable unit prices.

The reservation-based price is based on the type of services used by the client, which involves deciding the unit price for reserving the services, and the length of the reservation period. Final payment for the reservation is the product of the unit price for the reservation and the length of the reservation period.

Technical Support-related Pricing

Vendors provide different levels of technical support and charge clients by service

type and technical support time. IaaS clients provide great flexibility in service types. In contrast, integrative cloud service clients have fewer choices. However, the situation is reversed with respect to technical support. This could be due to the relative simplicity of IaaS and PaaS services. Clients can terminate the services anytime without incurring high costs. In contrast, SaaS services typically contain functions that are out of the client's control. More technical support from the vendors is needed when problems occur. In general, the list of support provided by IaaS vendors is shorter than that by integrative Cloud vendors. The only exception is Amazon, which has been expanding its line of technical support in recent years to include four types. (See Appendix A1.)

The support-related pricing is affected by the type of support services a client chooses and the length of the period that the support services are needed. It is a product of the support charge rate and the length of the support period.

Penalty-related Pricing

Services are experience goods: their tangible features do not fully reveal their true value. Software outsourcing contracts have a similar issue due to information asymmetry (Dey et al. 2010). Enhancing the completeness of the contract can help overcome this problem, but at a high cost (Hart and Moore 1999). In the practice of software outsourcing contracting, most vendors specify the penalties applicable when delivery is delayed (Whang 1992). Clients also have the right to terminate the contract.

In cloud services, service level agreements (SLA) serve as an incomplete contract. Service uptime guarantees are often provided in an SLA. A typical SLA includes terms specifying service characteristics and an uptime guarantee of 99.9%. Corre-

sponding penalty terms are defined in the SLA too. More often IaaS and PaaS offerings provide SLAs that include both uptime guarantee and penalty terms. But only a small part of SaaS offerings offer SLAs, such as Google's App for Business.

There appear to be two major types of penalty terms. Vendors such as Amazon and Microsoft Azure provide monetary compensation in the form of service credits, while some others, such as Google, provide free usage for a certain number of days in case service failure occurs. (See Appendix A2.)

2.5 Discussion

Quantity discounts are common in pricing strategy that incentivizes buyers to purchase greater than usual quantity. Research has shown that second-degree price discrimination, a nonlinear pricing strategy such as quantity discounts, is an effective way for vendors to segment clients, gain market power and obtain higher profits (Goldman et al. 1984, Monahan 1984). In the current cloud computing services market, only storage service vendors provide quantity discounts in the form of ladder-shaped tariffs. (See Appendix A1.) They offer clients who use the services bigger discounts on the unit prices. Other than that, quantity discounts are rarely used in any other categories of cloud services. For example, for an Amazon EC2 On-Demand Standard Instance (small) running on Linux/Unix, the pricing is fixed at $0.06 per instance-hour. There is no unit price difference for a customer who runs 10 instance-hours versus one who runs 10,000 instance-hours. For information goods, past research indicates that usage-based pricing with a quantity discount strategy is optimal when there are no transaction costs (Maskin and Riley 1984). Thus, it is viable for cloud vendors to include quantity discounts in pricing their services to incentivize cli-

ents to consume more services.

All the vendors reviewed in this study, except for Salesforce, provide uptime guarantees. Among these uptime guarantees, IaaS vendors do better by offering different uptime guarantees for different types of services. For example, Amazon provides a 99.9% uptime guarantee for S3, and a 99.95% uptime guarantee for EC2. Rackspace provides a 99.9% uptime guarantee for storage services and a 100% uptime guarantee for network availability.

Most of the SLAs include uniform penalties that the vendor must pay to all sorts of clients. However, the clients' attitude toward the risk of services downtime differs across applications and periods. Mission-critical enterprise applications carry a high cost for service downtime (Hiles 2005). In order to meet the diverse expectations, the vendor may wish to consider including customized penalty terms. They may outperform uniform penalty setting. It can be mutually beneficial to provide functions for negotiating penalty setting to satisfy different types of clients.

Furthermore, current cloud vendors provide a guarantee of short response time, ranging from hours to days, in the event of services downtime. This is different from a services uptime guarantee, which always includes penalty terms. There are no penalty terms that cover the conditions when the promised response time is not met. In case vendors do not fulfill their promised response time, clients have no contractual protection against the breach.

2.6 Concluding Remarks

According to the observations made by examining pricing schemes in the cloud computing services market, most components that are involved in the provision of

computing services could be charged separately. This is a real challenge for cloud services vendors though. On the one hand, overly complicated pricing structures make potential clients step back when they realize the complexity and unpredictable costs due to their lack of awareness (Perry 2010). On the other hand, an overly simple pricing model cannot meet clients' different needs and expectations.

This chapter presented a market survey and an analysis in order to shed light on important factors that have an impact on cloud services pricing. The observations and findings enable a richer understanding of vendors and clients perspective to supprt the conduct of more rigorous analysis for cloud computing services pricing.

3 Fixed-Price and Spot-Price Cloud Computing Services: A Damaged Services Perspective

3.1 Introduction

In this chapter, two different pricing mechanisms are analyzed for a cloud services vendor: fixed-price services, which are not subject to services interruption, and spot-price services that are interruptible.

Cloud computing services vendors deliver IT resources and software applications via the Internet. They can be scaled up and down to accommodate fluctuations in client demand. The market for cloud computing services has grown rapidly over the past decade. According to 451 Research (2013), cloud computing market revenue will grow at a compound rate of 36% and reach US$20 billion by the end of 2016.

Most vendors have adopted the usage-based pricing model, in which clients' payments are directly tied to their usage of computing services. Cloud services are consumed in a way similar to electricity and water, and typically are charged by the minute, hour, or month. In recent years, some more innovative pricing schemes have been implemented. For example, in 2006, Amazon Elastic Compute Cloud (EC2) started to sell its services based on an hourly fixed price. In 2009, Amazon allowed clients to purchase contracts for services sold as reserved instances. Then in 2010, Amazon offered services as instances with an hourly spot price. Spot prices change over time and, most of the time, are lower than fixed prices for reserved services, thus clients have an economic incentive to use them. But spot prices may also rise to a higher level in certain circumstances (e.g., when the services vendor is facing a shortage of resources to back the provision of services as spot instances). When spot prices

rise in the presence of higher demand for its services, Amazon will interrupt any job that is running as a spot-priced instance whose resources are bid out at a higher price. In other words, the spot-price services are interruptible. Services interruption can cause severe consequences, resulting in disutility and financial losses on the client's side. An example can be found in Howard (2011) in which a software application company, whose major functionalities were running as spot-priced instances, was unable to serve its clients for over a week. The incident was caused by a sudden spot price spike in September 2011, and the spot price remained at a high level for several hours and kept the core units of the company offline. Clients, thus, must balance the benefits and risks associated with the use of spot-price cloud computing services.

Such interruptible spot-price services constitute a defective version of existing fixed-price reserved services. Clients perceive them as having a lower expected value compared to the interruption-free reserved services. Interestingly, the lower value is not associated with lower service costs on the vendor's side: spot-price services share the same IT infrastructure with fixed-price reserved services, and their provision and delivery incur the same costs for the vendor. Instead, the lower value associated with spot-price services comes from the fact that the service vendor will deliberately impose interruption risks on clients, and thus it makes the services less attractive.

For these reasons, spot-price services are *damaged services*. They are similar to damaged goods, where a vendor has purposely modified some features of an existing good to make a lower quality version (Deneckere and McAfee 1996). Though it does not involve any additional cost savings, the *damaged goods strategy* has been shown to be an effective way to segment the market and conduct price discrimination. Will

the cloud vendor be able to do the same by damaging its services? It is an interesting research question that has not been investigated before. This work tries to seek answers for it. Furthermore, a vendor can offer fixed-price reserved services or spot-price on-demand services, or permit clients to use both. So it is interesting to ask: Should a vendor adopt a *hybrid pricing strategy*? What prices should it set? What level of consumer surplus can be obtained? How does social welfare change as the prices change? This work develops a model, and an analysis for the price and quality of cloud computing services offered with hybrid pricing. Based on the findings of an analytical model, I offer new results, and discuss some managerial implications.

In particular, this work studies a cloud marketplace with a monopoly services vendor and many potential clients. The vendor considers what to offer and how to charge for its services. There are multiple choices: the vendor could offer fixed-price, interruption-free reserved services only, or damage the services by introducing some level of interruption risk and offer spot-price on-demand services only. Or the vendor could use a *hybrid pricing strategy* in which both types of services are made available to the market. Potential clients have different levels of demand for computing resources. They will choose the services that give higher expected utility.

I am interested in studying hybrid pricing strategy and will address the following research questions. When is it optimal for the vendor to use a hybrid pricing strategy? What is the appropriate service interruption level for the damaged spot-price services? And how are consumers affected by the use of such a hybrid pricing strategy? This work offers insights that could help cloud vendors choose appropriate pricing strategy. First, it shows that offering damaged spot-price services only is always un-

desirable. In other words, offering fixed-price reserved services alone is more profitable to the vendor compared to offering spot-price services alone. It is an intuitive finding: with no cost reduction involved, the vendor should not damage its services and lower its quality from its clients' viewpoint. Second, the hybrid pricing strategy is not always appropriate for the vendor. Sometimes it could be worse than if it offers fixed-price reserved service alone. Whether the vendor should employ the hybrid pricing strategy depends on two factors: the price levels for spot services and the clients' sensitivity levels to services interruption.

When the expected spot price is high, or clients are highly sensitive to services interruption, the hybrid pricing strategy is profit-maximizing. The use of the hybrid pricing strategy, in many cases, will lead to larger market coverage and higher total social welfare, but lower consumer surplus. A more interesting result is that, under the hybrid pricing strategy, the vendor's profit will be higher when the spot-price services are subject to higher interruption risks. This suggests that the vendor should deliberately damage its spot-price services by attaching a high level of services interruption risk to them, so that they will not compete with the fixed-price reserved services too severely and hurt the vendor.

The rest of this chapter is organized as follows. Section 3.1 reviews related literature. Section 3.2 presents the model setting. Section 3.3 lays out the analyses and major findings. Section 3.4 discusses some model extensions and possible impacts on the results. Section 3.5 explores the managerial implications of this research and Section 3.6 concludes.

3.2 Literature

This work is built on two streams of research: pricing goods and services with uncertain demand, and quality differentiation. Three other areas of inquiry on pricing goods and services with uncertain demand are related to this work. One is research on *information goods*, the second is on *peak-load pricing*, and the third is on *revenue yield management*.

The research on pricing information goods has analyzed business models with fixed fees and usage-based fees. Sridhar et al. (2009) developed a model in which demand was uncertain in clients' usage frequency and utility-per-use of the goods and services, and showed that a monopolist should employ usage-based pricing when transaction costs are low. With competition though, fixed pricing often outperforms usage-based pricing. This result is also supported by other research on selling electronic goods with fixed subscription fees or with pay-per-use fees (Fishburn et al. 1997, Fishburn and Odlyzko 1999). These works all considered zero marginal production costs and positive transaction costs associated with pay-per-use pricing. Sundararajan (2004) reported that a monopolist using both performed no worse in the presence of positive transaction costs—and sometimes better—than when only using usage-based pricing. The setting in this research is similar, but spot prices are different from usage-based prices. Usage-based pricing involves a fixed payment per unit. In the current setting, spot prices vary over time and involve unexpected services interruption. This novel usage-based pricing has not been studied in the IT services-related literature yet.

The second line of work deals with *peak-load pricing* for non-storable products such as electric power, where price discrimination is appropriate, and linear (Steiner

1957) or convex (Boiteux 1960) costs have been assumed. Understanding the cost structure, the relationship between marginal production cost and the capacity of the plant, is critical (Wilson 1972). However, this study does not consider the problem of capacity planning for cloud vendors. Nevertheless, it will consider a special condition under which the vendor can use capacity as a tool to improve its profit.

The third line of related research is *revenue yield management*, in which pricing—among several other aspects such as inventory control and overbooking—is an important aspect (McGill and Van Ryzin 1999). Gallego and Van Ryzin (1997) showed that, when studying the revenue management for airline tickets, there is a natural duality between ticket prices and seat-allocation decisions. In light of this duality, Li (1994) proved that it is optimal to offer a small number of fare classes, with different restrictions, such as introducing non-refundable and no-luggage conditions. In addition, the airlines have used spot prices with infrequent changes, and leveraged high and low prices to ration capacity (Dana 1999). Different from these previous works, for Amazon.com's EC2 services, hourly spot-price changes have been used to support value-based resource allocation.

Another related stream of research includes studies on *quality differentiation*. Quality differentiation enables the vendors to segment the market and price discriminate. The structure of large sunk costs and low variable costs (for producing the product of different qualities) makes versioning suitable for value-based pricing and quality differentiation. Prior studies on information goods suggest a limited number of versions of the information goods with different quality levels should be offered. Varian (1997) suggested offering different versions when clients have different preferences

for quality, or their preferences are hard to observe. A common rule of thumb is to offer only two versions: a high quality version and a low quality version. This applies to the settings in this work: it studies whether a cloud services vendor should offer the services with different quality levels at different market prices. Some other works on damaged goods are related. It has been well-documented how manufacturers can improve their profits by intentionally damaging a portion of their goods and then employing price discrimination (Deneckere and McAfee 1996). The improvement in profit, however, does not depend on the distribution of client valuations, but instead on the value of the damaged goods relative to the undamaged ones (McAfee 2007).

3.3 The Model

Consider a monopoly vendor that offers fixed-price reserved services and spot-price on-demand services. To use the reserved services, clients must buy a *reserved services contract.* They pay the *reserved services price* $P_{Reserved}$ in advance to reserve N units of resources. (See Table 3.1 for modeling notation.)

Table 3.1 Model Variables and Parameters

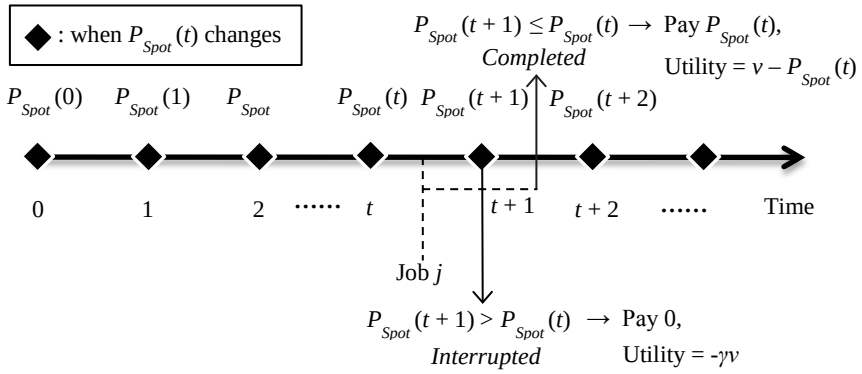
VARIABLE	DEFINITION
P	Prices for a cloud computing services contract (with *Reserved* or *Hybrid* subscripts)
$N_{Reserved}$, N_{it}	Resource capacity limit of reserved services contract, the number of client *i*'s reserved services units left at the beginning of period *t*
$P_{Spot}(t)$, $\bar{P}_{Spot}$; P_L, P_H	Spot price (at time *t*) and expected spot price; low and high spot prices
$Prob_L$, $Prob_H$	Probability of low or high spot prices ($Prob_H$ = 1 - $Prob_L$)
$Prob_{Interruption}$	Probability of spot-price services interruption, $Prob_{Interruption}$ = $Prob_L$(1 - $Prob_L$)
v, $\bar{v}$, v^0; v_L, v_H	Single job value, expected value, threshold job value of submission strategy in the limited reserved services capacity case; lower bound, upper bound on job value
λ_i, λ_{it}^+, Λ	Job arrival rate of client *I*; job arrival rate of client *i* in the other periods after period *t*; maximum job arrival rate
$U_{Reserved}$, U_{Spot}	A client's utility from jobs run with fixed-price reserved services; a client's utility from jobs run with spot-price on-demand services
γ	Clients' sensitivity to services interruption
$\tilde{v}$	Expected job utility difference for reserved and spot-price services
π	Vendor's profit (with *Reserved, Spot or Hybrid* subscripts)
CS, SW	Consumer surplus; social welfare (with *Reserved, Spot* or *Hybrid* subscripts)

To use the spot-price services, a client will pay the current spot price. I examined spot price data from Amazon EC2 from January to July 2012, and found that spot prices were at the base level much of the time, but fluctuated between the base level and a higher price for the rest of the time. A model with spot prices as a variable with two values is appropriate: a low price P_L with a probability of $Prob_L$, and a high price P_H with a probability of $Prob_H$ = 1 - $Prob_L$. Here, P_L, P_H, $Prob_L$, and $Prob_H$ are common knowledge to the vendor and its clients.

The model covers multiple periods. (See Figure 3.1.) At the beginning of each, the spot price $P_{Spot}(t)$ will change and be announced to all clients. Such services can

be interrupted when capacity is limited and the spot price rises. Consider job j. At time t, the spot price $P_{Spot}(t)$ is announced by the vendor. A client submits job j during time t and $t + 1$ for spot-price services. At time $t + 1$, the vendor revises the spot price to $P_{Spot}(t + 1)$ in the presence of new demand. If $P_{Spot}(t + 1) > P_{Spot}(t)$, job j will be interrupted by the vendor and the client will not pay anything. If $P_{Spot}(t + 1) \leq P_{Spot}(t)$, job j will execute to completion, and the client will pay $P_{Spot}(t)$. To simplify, each job execution will take one unit of time, so each job is subject to one spot price change.

Figure 3.1 Job executions and payment timeline



Clients have heterogeneous demand in terms of the number of jobs that need to be run. I only consider their jobs that will arrive in the time period described in Figure 3.1. Client i's jobs arrive following a Poisson distribution with the arrival rate λ_i, and are independent and identically distributed (i.i.d.) across all periods. I use the Poisson job arrival process because it has attractive theoretical properties yet well describes job arrivals in real world data center (Frost and Melamed 1994, Li et al. 2006), and it has been widely used by previous research (Li et al. 2006, Li et al. 2007, Parolini et al. 2010). Some clients will have more jobs. The job arrival rate λ_i is uniformly distributed over $[0, \Lambda]$, with Λ as the largest number of jobs that may arrive, and $F(\lambda_i)$ is the

cumulative distribution function. Each client knows its own λ_i, while the cloud computing services vendor only knows the distribution of λ.

The jobs have varied value to clients. This reflects the reality that a client may use cloud computing services for different purposes. The value of a single job follows a uniform distribution over the range (v_L, v_H). And all the clients have the same job-value distribution. If a spot-price job is interrupted, the client will incur disutility $-\gamma v$, proportional to the job value v, with $0 < \gamma < 1$. Here γ measures a client's sensitivity to services interruption. Let us further assume that $P_H < v_L$. This means that spot-price services are cheap, so most clients stay in the market.

3.4 Analysis and Results

Consider a market in which only fixed-price reserved services are offered, and then a market in which only spot-price on-demand services are offered by the vendor. My analysis will consider hybrid pricing, and will compare these three markets in terms of the vendor's profit, consumer surplus, and social welfare.

Vendor Offers Services with Single Pricing Scheme

A vendor's decision on pricing when it offers its cloud computing services with one type of pricing scheme alone affects its profit, and consumer surplus, and social welfare. I will discuss the case when the vendor only offers the fixed-price reserved services, and then the case when the vendor only offers spot-price on-demand services.

Fixed-Price, Non-Interruptible, Reserved Services Only. When the vendor offers its computing resources through reserved services only, its clients need to choose

whether to purchase a reserved services contract. The decision has to be made at the beginning of the *k*-period timeline (see Figure 3.1 for the timeline) at time 0. The optimal reserved services contract ($P^*_{Reserved}$, $N^*_{Reserved}$) can be derived with the proposed model.

Lemma 1, presented first, states that the vendor should not set any capacity limit in the reserved services contract, if it wishes to achieve optimality. Limiting resource capacity will reduce a client's valuation of the contract. The vendor will have to lower its price or tolerate a decrease in demand, and will lose profit as a result.

- **Lemma 1 (Capacity Limit of Optimal Reserved Services Contract).** *When a cloud vendor offers reserved services only, any finite limit of resource capacity N is not optimal.*

When the vendor offers unconstrained usage in a reserved services contract, a client will be able to have all its job arrivals executed on reserved services. The client whose expected total utility from job executions equal to the reserved services contract price $P_{Reserved}$ will be indifferent between purchasing the contract and staying out of the market. Let $\lambda^*_{Reserved}$ be the job arrival rate for this marginal client. Then $\lambda^*_{Reserved}\bar{v} = P_{Reserved}$, where $\bar{v}$ represents the expected job value. The vendor's profit function is $\pi_{Reserved} = (1 - F(\lambda^*_{Reserved}))P_{Reserved}$, and it will maximize profit by choosing an appropriate price.

- **Proposition 1 (Reserved Pricing Strategy).** *When the cloud vendor offers reserved services only, it should price the services contract at $P^*_{Reserved} = \bar{v}\Lambda / 2$ and not constrain the capacity limit to $N^*_{Reserved}$.*

The marginal client will have job arrival rate $\lambda^*_{Reserved} = \Lambda / 2$, the vendor's profit will be $\pi^*_{Reserved} = \bar{v}\Lambda / 4$, and consumer surplus and total social welfare will be $CS^*_{Reserved} = \bar{v}\Lambda / 8$ and $SW^*_{Reserved} = 3\bar{v}\Lambda / 8$. At the optimum, half of the market, including those clients with a high level of demand, will purchase the reserved services contract. The vendor will receive a major portion of the total social welfare, as its profit is two times the consumer surplus.

Spot-Price, Interruptible, On-Demand Services Only. Consider a market where a vendor offers only spot-price services. Spot prices are assumed to be exogenously determined. The vendor controls the price change frequency: it decides the probability of a low price (P_L) and a high price (P_H) to appear. Thus, it can maximize profit via an optimal valued $Prob_L$, the probability of a low spot price to occur.

The vendor, however, cannot set the probability of P_L and P_H too low. Consider a job that arrives during $[t, t + 1]$. When it is submitted, the spot services price is $P_{Spot}(t)$; and during its execution the spot price will be revised to $P_{Spot}(t + 1)$. If $P_{Spot}(t) = P_H$, the service interruption probability for this job is zero and the expected net utility for the client is $v - P_H$. If $P_{Spot}(t) = P_L$, the probability of service interruption is $1\text{-}Prob_L$. The expected net utility for the client then becomes $(v - P_L)\, Prob_L - \gamma v\, (1 - Prob_L) = v\, (Prob_L + \gamma\, Prob_L - \gamma) - P_L\, Prob_L$. Note that when $Prob_L$, the probability of P_L to occur, is very small, $Prob_L < v_H \gamma / [v_H (1 + \gamma) - P_L]$, the vendor will not be able to operate the spot-price services. It is likely that a job runs with spot-price services will be interrupted when its price is P_L, so the client's net utility will always be negative. As a result, no client will submit a job for spot-price services when P_L is observed. In addition, when $v_H \gamma / [v_H (1 + \gamma) - P_L] \leq Prob_L < v_L \gamma / [v_L (1 + \gamma) - P_L]$, only jobs with high

value will yield positive expected net utility. In order to make the analysis more comprehensive and interesting, let us consider a scenario in which clients will submit all their jobs for spot-price services. Hereafter, this condition $Prob_L \geq v_L \gamma / [v_L (1 + \gamma) - P_L]$ is assumed to hold.

Since all clients will use spot services, the demand for spot services is $\Lambda / 2$. The vendor's profit will be $\pi_{Spot} = (P_L\ Prob_L^2 - P_H\ Prob_L + P_H)\ \Lambda / 2$. The vendor's decision is to set the probability of a low spot price $Prob_L^*$. This solution turns out to be dependent on the ratio of P_H to P_L, as shown by Proposition 2.

- **Proposition 2 (Spot Pricing Strategy).** *When a cloud vendor offers spot services only, it should set $Prob_L^*$ as follows:*

  Case (1) *when* $P_H \geq P_L(1 + \frac{\gamma v_L}{(1+\gamma)\ v_L - P_L})$, $Prob_L^* = \frac{\gamma v_L}{(1+\gamma)\ v_L - P_L}$; *and*

  Case (2) *when* $P_H < P_L(1 + \frac{\gamma v_L}{(1+\gamma)\ v_L - P_L})$, $Prob_L^* \rightarrow 1$.

In Case (1), when the difference in spot prices is large, the probability of services interruption should be decreasing in γ. In the extreme case of $\gamma \rightarrow 0$, meaning that a client incurs almost zero disutility from interruption, the vendor will not need to offer spot-price services. In Case (2), when the difference in spot prices is small, the vendor's best action is to set $Prob_L^*$ equal to 1, so any service interruption risk vanishes.

The Spot Price Strategy Proposition (P2) shows that how often the vendor should make an upward price adjustment depends on the relative magnitude of P_H and P_L, and is moderated by the client's sensitivity to services interruption γ. There are two cases. Which case the vendor will face is determined by the outcome of the comparison between the high spot price P_H and a compound term including the low spot price

$P_L\,(1 + \frac{\gamma v_L}{(1+\gamma)\,v_L - P_L})$. The latter term is increasing in the client's sensitivity to services interruption γ. Therefore, the result of the comparison will possibly be different at different levels of γ. That means the vendor may face Case (1) as described in Proposition 2 when γ is small, and will face Case (2) when γ is big.

The vendor's profit will also be different in the two cases. The following corollary shows this finding.

Corollary 1. *The vendor will gain a higher profit when it is able to follow Case (1) described in Proposition 2.*

In addition, in Case (1), consumer surplus will be much lower compared to Case (2), due to a higher level of services interruption risk introduced by the vendor. As a result, social welfare will also be lower.

Vendor Offers Fixed-Price and Spot-Price Services: Hybrid Pricing Strategy

Pricing strategy. Hybrid pricing may cause two opposite effects. Spot-price services will serve clients who stay out of the market before the services become available. This will increase the usage of a vendor's infrastructure and generate more profit for the vendor. Spot-price services may cannibalize the fixed-price services though. Some clients who would have purchased reserved services may now opt for spot-price services. So the effect of offering spot-price services in addition to existing fixed-price cloud computing services is unclear.

The indifferent client will obtain the same expected utility from using reserved and spot-price services. This client is identified via the job arrival rate: $\lambda^*_{Hybrid} = P_{Reserved} / (\bar{v}\,(1 + \gamma)\,Prob_L\,Prob_H + P_L\,Prob_L^2 + P_H\,Prob_H)$. For simplicity, let us introduce $\tilde{v} = \bar{v}(1 + \gamma)\,Prob_L\,Prob_H + P_L\,Prob_L^2 + P_H\,Prob_H$, and hence $\lambda^*_{Hybrid} = P_{Reserved} / \tilde{v}$.

The vendor then needs to maximize profit π_{Hybrid} = (1 - $\frac{2\tilde{v} - \bar{P}_{Spot}}{2\tilde{v}^2 \Lambda} P_{Reserved}$) $P_{Reserved}$. The vendor will decide: (1) the optimal price for the fixed-price reserved services contract (P^*_{Hybrid}), and (2) the probability that low spot prices will occur ($Prob^*_L$). Solving the vendor's optimization problem leads to:

- **Proposition 3 (Hybrid Pricing Strategy).** *To implement the hybrid pricing strategy, the vendor should price a reserved services contract at* P^*_{Hybrid} = $\tilde{v}^2 \Lambda$ / (2$\tilde{v}$ - $\bar{P}_{Spot}$)*. In addition, the vendor should always configure the optimal value of* $Prob^*_L$ *to be smaller than 0.5* and $Prob^*_L$ *is increasing in client sensitivity to service interruption* γ*.*

As a result, clients with job arrival rates greater than λ^*_{Hybrid} = $\tilde{v}\Lambda$ / (2$\tilde{v}$ - $\bar{P}_{Spot}$) will purchase fixed-price reserved services, while the rest will use spot-price on-demand services. The vendor's profit will be π^*_{Hybrid} = $\tilde{v}^2 \Lambda$ / (4$\tilde{v}$ - 2$\bar{P}_{Spot}$).

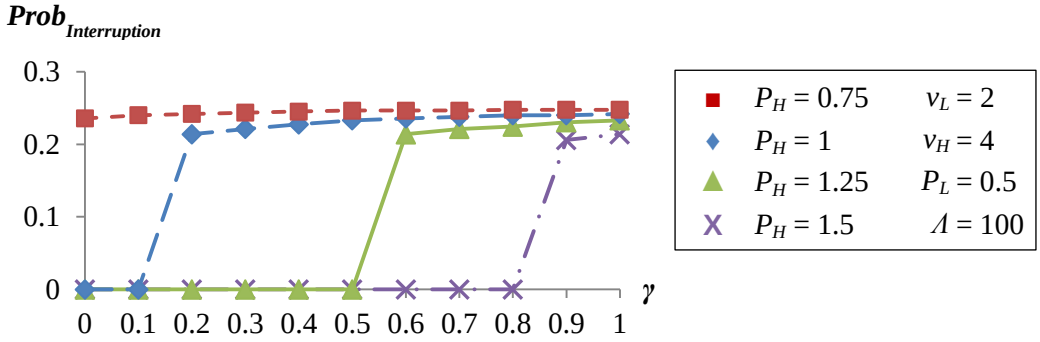
Several findings are salient. First, due to the cannibalization effects, fewer clients will purchase the fixed-price reserved services contract. Second, the market share of fixed-price services is always smaller than that of spot-price services. This means that the market will have more clients using spot services than reserved services. Third, with hybrid pricing, the vendor's profit will come from both fixed-price and spot-price services. Though spot-price on-demand services will attract the majority of the clients, which services contribute more to the vendor's total profit will depend on clients' sensitivity to services interruption. If clients incur high disutility due to services interruptions (e.g., when $\gamma > 1 - \frac{\bar{P}_S}{2\bar{v} Prob_L Prob_H}$), fixed-price reserved services will yield higher profit for the vendor than spot-price on-demand services. If clients incur low

disutility from services interruptions (e.g., when $\gamma < 1 - \frac{\bar{P}_S}{2\,\bar{v}\,Prob_L Prob_H}$), spot-price services will contribute more. So, the interruptible and uninterruptible services will not be equally profitable to the vendor, except when $\gamma = 1 - \frac{\bar{P}_S}{2\,\bar{v}\,Prob_L Prob_H}$.

Finally, the vendor should increase the reserved services price and the probability of a low spot-price level in the market when clients are increasingly sensitive to services interruptions as γ increases. This will improve the vendor's profit. The underlying rationale is as follows. Spot-price services with interruption risks are inferior. Making such services subject to higher interruption risks increases the value difference between the services. As a result, the vendor is able to raise its price for reserved services and can segment the market based on client self-selection behavior.

The effect of γ on $Prob_{Interruption}$ is moderated by the spot-price ratio P_H / P_L. (See Figure 3.2.) Based on the optimal value of $Prob_L^*$, interruption risk $Prob_{Interruption}$ increases as client sensitivity to services interruption γ increases. On the other hand, at the same level of γ, say $\gamma = 0.9$, when P_H goes from 0.75 to 1.5, interruption risk decreases.

Figure 3.2 Probability of services interruption as a function of interruption risk sensitivity



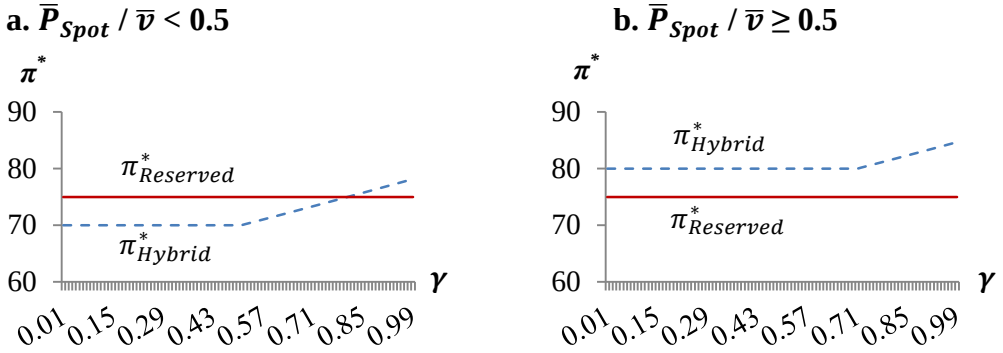
Impact of spot-price interruptible services on fixed-price reserved services. Next will be the comparison of the vendor's profit under different strategies. I have proved that, with a one-price strategy, in general, it will make sense for the vendor to offer fixed-priced reserved services, since they are more profitable than spot-price on-demand services. With the hybrid pricing strategy, however, profit may or may not increase.

- **Proposition 4 (Impact of Spot-Price On-Demand Services).** *A sufficient condition for the vendor to increase its profit by introducing spot-price services in addition to the existing fixed-price reserved services is* $\bar{P}_{Spot} / \bar{v} \geq 0.5$, *where* $\bar{P}_{Spot}$ *is the expected spot price,* $\bar{v}$ *is the expected job value. The greater the clients' sensitivity to services interruption risk is, the higher profit the vendor obtains.*

This proposition reveals a key finding: the hybrid pricing strategy may be beneficial or hazardous to the vendor. The spot prices (P_L and P_H) and clients' sensitivity to services interruption (γ) determine how the vendor's profit changes. When clients are sensitive to services interruption, the vendor is likely to gain a profit increase by offering spot-price services in addition to fixed-price reserved services.

This finding may seem counter-intuitive, but there is some rationale for it. Recall that there is a cannibalization effect when introducing spot-price services to market with existing fixed-price reserved services. It will improve the vendor's profit only if the spot-price services are not too attractive: clients will need to bear high disutility once jobs are interrupted. (See Figure 3.3.)

Figure 3.3 The vendor's profit versus interruption risk sensitivity



Note: The simulation values in the left-side figure are P_L = 1.3, P_H = 1.5. In the right-side figure, they are P_L = 1.4, P_H = 1.6; $\bar{v}$ = 3. All the other parameters are the same in both figures.

The main difference between the parameter settings in Figure 3.3 (a) and Figure 3.3 (b) is in the relationship between the expected spot price $\bar{P}_{Spot}$ and the expected job value $\bar{v}$. The example in Figure 3.3 (a) indicates that, when the condition $\bar{P}_{Spot} / \bar{v} \geq 0.5$ does not hold, the hybrid pricing strategy could improve the vendor's profit only when clients are very sensitive to services interruption. If $\bar{P}_{Spot} / \bar{v} \geq 0.5$ holds, as shown in Figure 3.3 (b), the vendor will always achieve higher profit by offering spot services in addition to reserved services.

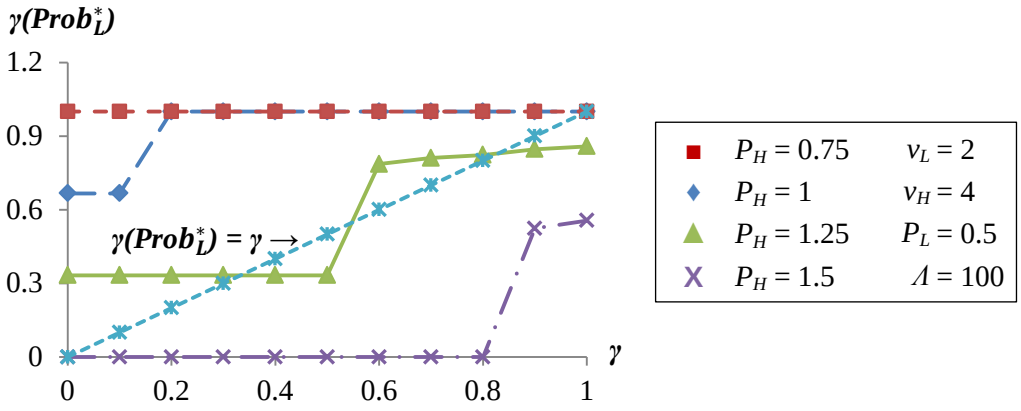
Consumer surplus and social welfare. Adding spot-price on-demand services to existing fixed-price reserved services always leads to a market expansion. This may not result in a higher overall consumer surplus or social welfare though. So we can conclude:

- **Proposition 5 (Consumer Surplus and Social Welfare).** *A sufficient condition under which adding spot-price on-demand services will increase consumer surplus is: $\gamma < (1 / 2 - 2\bar{P}_S / \bar{v})$. When either of the two conditions holds, (1) $P_H / \bar{v} < 0.943$ or (2) $\gamma < 0.939$, adding spot-price on-demand services will in-*

crease social welfare.

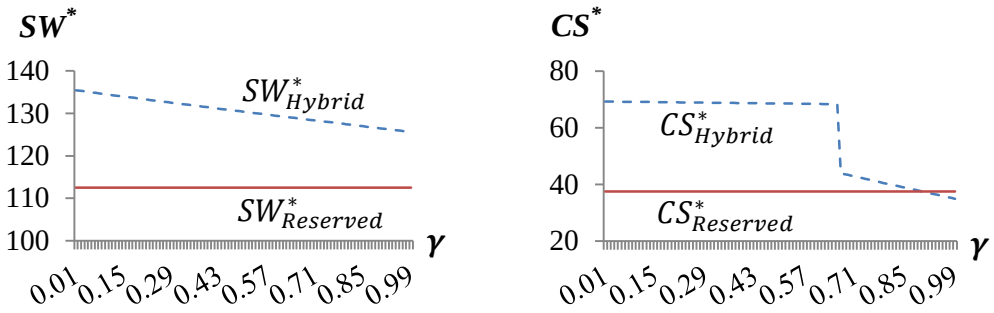
The condition for increasing consumer surplus will not be easily met when the high spot price P_H is much higher than the low spot price P_L. Building on the prior numerical analysis, a critical value $(2 - 4(P_L\, Prob_L^2 + P_H\, Prob_H) / \bar{v})$ can be calculated and compared with the clients' sensitivity to services interruption γ. When P_H is much higher than P_L, the clients' sensitivity to services interruption is almost always higher than the calculated critical value. (See Figure 3.4 when $P_H = 1.5$.) When the high spot price is much higher than the low spot price, consumer surplus is likely to decrease.

Figure 3.4 Critical value of the clients' sensitivity γ to interruption risk based on the optimal probability of a low spot price to occur



On the other hand, when the high spot price is close to the low spot price (see Figure 3.4 when $P_H = 0.75$) clients will achieve higher consumer surplus for most of the time. This is indicated by the two dashed lines with square and diamond marks at the top. (See Figure 3.4.)

Figure 3.5 Consumer surplus and social welfare comparisons



Note: The assumed values in the figure are: $v_L = 2$, $v_H = 4$, $P_L = 1.4$, $P_H = 1.6$, $\Lambda = 100$.

Adding spot-price on-demand services will expand the market size for the vendor, so more clients will be served, which has the potential to increase social welfare. It is not straightforward to determine whether social welfare will increase or decrease though. The vendor will be able to sell fixed-price reserved services to fewer clients, compared to when spot-price on-demand services are not offered. With fewer cloud computing jobs executed as either spot services or reserved services, there is a greater possibility that they will be completed, resulting in full job utility for clients. The profit decrease in reserved services, however, will be much smaller when clients are highly sensitive to services interruptions. Plus, with the additional jobs that get completed with spot-price on-demand services, the overall level of social welfare may increase. But this will occur only when the clients' sensitivity (γ) exceeds the critical value given in the Consumer Surplus and Social Welfare Proposition (P5).

3.5 Model Extensions

To this point in the chapter, the focus has been on fixed-price, no capacity limit services ($N = \infty$), for the case with reserved services only, which gives clients a price discount. A client pays the price $P_{Reserved}$ and can use as many computing resources as

it needs. So clients with higher demand pay, on average, a lower price for each service unit they consume than clients with lower demand. The price discount is only limited by the client's demand. It means that the higher a client's demand, the more discount will the client get from using the fixed-price reserved services.

If the vendor introduces a services capacity limit on a reserved services contract, there will be two possible effects. First, the reserved services will become less attractive. As a result, probably fewer clients will buy it. Second, clients with high demand will find the reserved services contract to be insufficient to handle all their jobs. They will submit some of the jobs for spot-price on-demand services. Furthermore, since jobs are heterogeneous in their value, high-value jobs will be more likely to be submitted for reserved services. The first effect will potentially cause the vendor to lose profit, while the second effect may increase it. The overall effect on the vendor's profit of introducing services capacity limit is unclear.

It is common in practice that a service vendor will limit the capacity that the client can access and consume. I will next consider this case, when the vendor puts a capacity limit on its reserved services contracts. The term *capacity* refers to the usage upper limit associated with a reserved services contract.

Clients' Decision for Job Submission

Consider a client's job submission decision when the vendor uses a hybrid pricing strategy and assigns a capacity limit to the reserved services contract. When there is no capacity limit, clients who purchased the reserved services contract will submit all their jobs for reserved services. But their job submission behavior will change if capacity is limited. Using reserved services can secure the completion of a job for the

client. But using spot-price services will force the client to bear the risk of services interruption and potential disutility. In addition, when reserved resources are limited and job value varies, the client will try to use the limited reserved services to secure as much value as possible. Clients with high demand will submit some jobs for spot-price services and save reserved resources for future use. The impact of this job submission strategy and how the vendor's profit will be affected are worthy of further study.

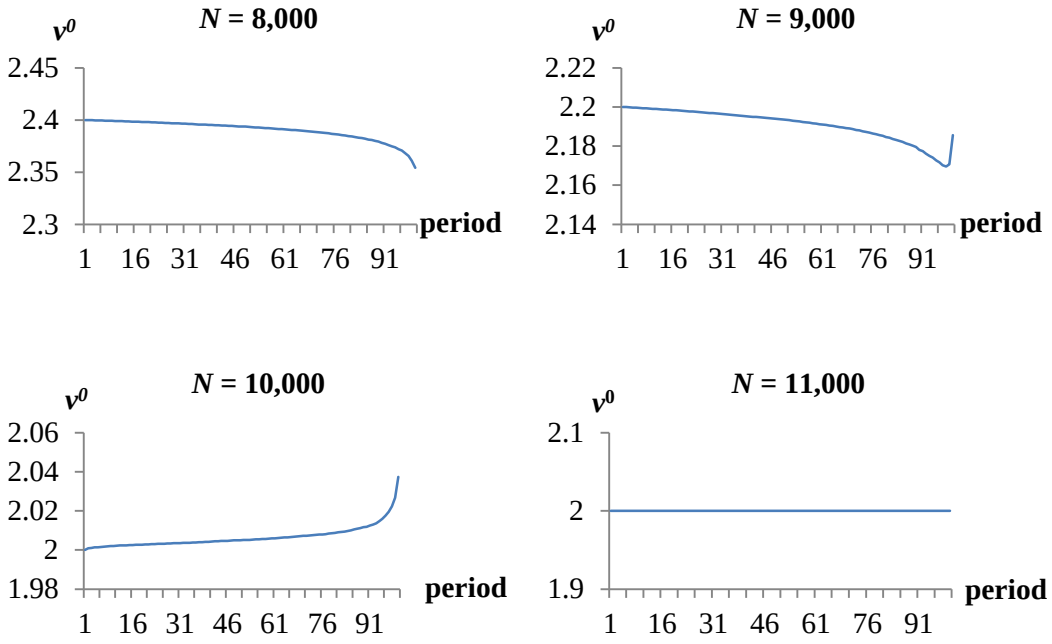
Consider one simple example. When two jobs arrive, one with a high value and the other with a low value, and when the reserved services contracts are limited to five units and there are four periods in total, using reserved services to complete the high-value job and using spot-price services to execute the low-value job is an optimal strategy. My choice of five units and four periods is for illustration: it does not mean that clients will change their job submission behavior because of the limit of this specific number of service units or number of periods. It is to show that, although a client still has enough reserved resources for all job executions in certain period, it may save some for high-value jobs that will arrive in the future. The principle is to only let low-value jobs face the risk of services interruption. The same principle applies when there are more than two arriving jobs and insufficient reserved resources.

When the reserved resources are not enough, the client needs to decide which jobs to submit for reserved services, and which jobs to submit for spot-price services. According to the principle I just discussed, the client needs to decide a threshold value v^0 that separates arriving jobs into two groups: a high-value group that will go to reserved services, and a low value group that will go to spot-price services.

The threshold job value v^0, however, has to be revised in each period, due to the fact that clients are uncertain about the exact number of jobs that will arrive in the future. At the beginning of period *t*, client *i* will know how many jobs have arrived and have been executed and how many units of reserved services are left. It will also update the distribution of future job arrivals, which follows a Poisson distribution with job arrival rate $\lambda_{it}^{+} = \left\lfloor \frac{\lambda_i}{k} \right\rfloor \times (k - t + 1)$. Let N_{it} be the number of the client's reserved services units left at the beginning of period *t*. Then we can derive v^0 for client *i* at period *t*: $v_{it}^0 = v_L + (1 - \min(1, N_{it} / \lambda_{it}^{+}))(v_H - v_L)$.

The threshold job value (v^0) changes over time. To illustrate this, I simulated the job arrival process for a client with a job arrival rate $\lambda = 100$ in a 100-period timeline. The threshold job value will be updated at the beginning of each period before any job arrives. The capacity limit associated with a reserved services contract is set at four levels: *N* takes one of the four values in the set {8,000, 9,000, 10,000, 11,000}. The job arrival process over the 100 periods is simulated 10,000 times. The average threshold job value is calculated for each period. (See Figure 3.6.) When the capacity limit is less than overall job arrivals, as in the cases N =8,000 and N = 9,000 shown in Figure 3.6, clients will start to save reserved resources earlier for future use. When the capacity limit is equal to or greater than overall job arrivals, as the cases N = 11,000 and N = 11,000, clients will either start to save reserved resources later or not save them at all.

Figure 3.6 The dynamic threshold for job value

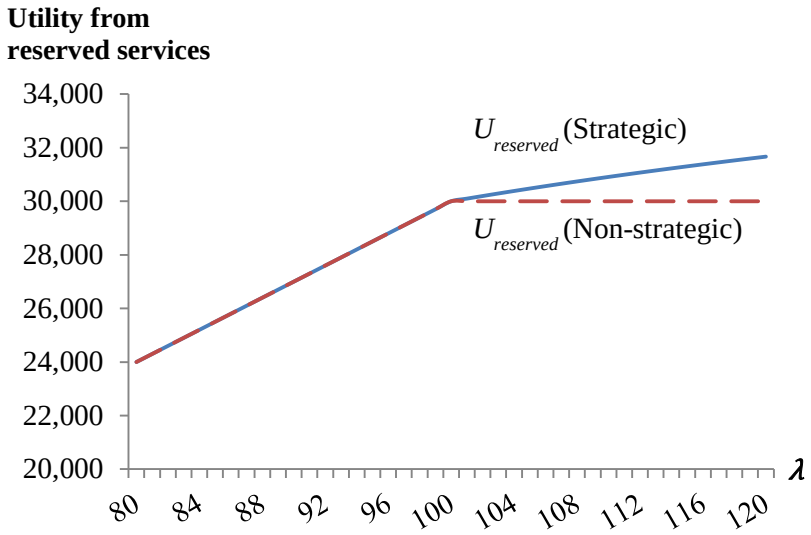


When the client with a job arrival rate λ is strategic in its job submission, it follows the threshold job value strategy, and it will achieve the utility from reserved services: $U_{Reserved} = \int_0^N \bar{v}x\, f_\lambda(x)dx + \int_N^\infty E(v|v > v^0)\, N\, f_\lambda(x)dx$. The first term is the utility when the total number of jobs is less than the capacity limit, and the second term presents the utility when total number of jobs exceeds the capacity limit. The utility function indicates there will be a knot in the utility curve as the job arrival rate increases. Figure 3.7 depicts $U_{Reserved}$ for different values of λ, with a capacity limit N = 10,000. Strategic clients with high demand achieve higher utility than non-strategic clients when λ is greater than 100.

Figure 3.7 indicates that strategic clients with high demand may be willing to pay a higher price. The utility curve for such clients is flatter when the job arrival rate is greater than the capacity limit. The total utility of a client represents its highest will-

ingness-to-pay for the service. The flatter slope, however, indicates that when clients are strategic in job submission, they are more sensitive to changes in the price of the reserved services contracts when their job arrival rate exceeds capacity limit N than when their job arrival rates are less than capacity limit N.

Figure 3.7 The utility $U_{Reserved}$ of strategic and non-strategic clients with different job arrival rates λ



Note: The parameter values are: $v_L = 2$, $v_H = 4$, $N = 10000$, and $k = 100$.

Reserved Services Contracts with Finite Resource Capacity

Now I will show how the vendor's profit will be affected by the capacity limit associated with the reserved services contract, and discuss how the vendor's pricing decisions may change in this case.

The first observation is that imposing the capacity limit in the reserved services contract but keeping the contract price unchanged will cause a vendor to lose profit from reserved services but gain profit from spot-price services. With the capacity limit, if a client still submits all its jobs to reserved services before the capacity limit is

reached, its net utility will be $U_{Reserved} = \int_0^N \bar{v}x\, f_\lambda(x)dx + \int_N^\infty \bar{v}N\, f_\lambda(x)dx - P_{Reserved}$. The first term is the expected total utility gained when the number of total arrived jobs is less than the capacity limit N, and the second term represents the expected total utility when the number of total arrived jobs exceeds the capacity limit N. When job arrivals exceed the capacity limit, the utility is bounded at $\bar{v}N$. As a result, the overall net utility will be smaller when capacity is limited than when capacity is not limited.

In addition, even if the client is strategic in job submissions, its overall expected utility will still be lower. While the total utility from using spot-price services for all job arrivals is not affected by the capacity limit, the client who, in the unlimited capacity case, is indifferent between purchasing the reserved services contract and purchasing spot services will find it no longer worth buying now. Because $U_{Reserved}$ increases in the client's demand (based on the job arrival rate λ), the marginal client will have higher demand. Consequently, the vendor will lose some clients for reserved services, which will lead to a lower profit. On the other hand, job arrivals that exceed the capacity limit will be processed by spot-price services, which will yield additional profit for the vendor. The overall effect of the limited capacity of the reserved services contract will depend on the magnitude of the capacity limit.

A second observation is that limiting the resource capacity associated with a reserved services contract will make the overall spot services usage increase. A client with high demand will submit some jobs for spot-price services even if it has already purchased the reserved services contract, because the capacity limit may not be sufficient for all of its jobs. In addition, some clients will refrain from buying reserved services due to the capacity limit. As a result, overall spot services usage will increase.

Following the above discussions, along with the hybrid pricing strategy, we can see a potential for capacity to be used as a tool to further improve the vendor's profit. This is similar to the practice of rationing the availability of capacity that is sold at a lower price. This is known as yield management, which has been commonly practiced by airlines, hotels, and several other industries (Dana 1999). The cloud computing services vendor's problem now is whether it can limit the capacity associated with the reserved services contract to an appropriate level so that its profit will increase.

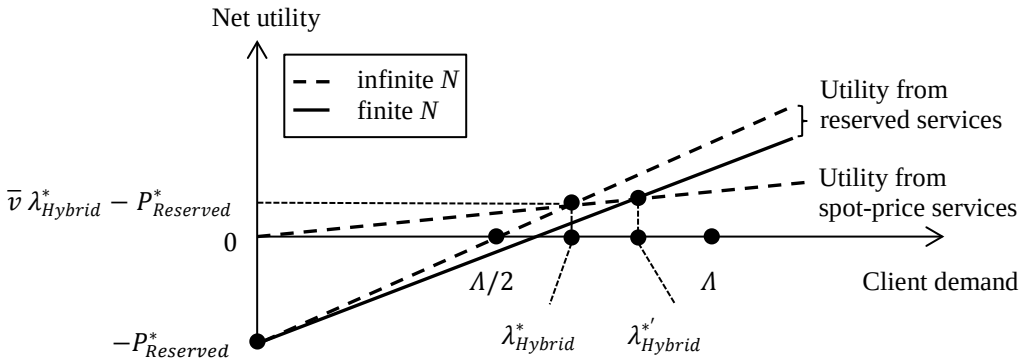
To find out the answer, I first need to re-examine the demand of the marginal client and find out to what extent the profit from reserved services will be affected. When there is a capacity limit, the job arrival rate of the marginal client must satisfy the condition: $\lambda'^{*}_{Hybrid} = arg_{\lambda}\{\int_0^N \bar{v}x\, f_{\lambda}(x)dx + \int_N^{\infty} \bar{v}N\, f_{\lambda}(x)dx - P_{Reserved} + \int_N^{\infty} U_{Spot}(x - N)\, f_{\lambda}(x)dx = \lambda U_{Spot}\}$. The first term in the bracket represents the utility when the number of total jobs is less than the capacity limit, the second term is the utility from jobs running with reserved services when the number of total jobs exceeds the capacity limit, and the fourth term represents the net utility from jobs that cannot run with reserved services due to the capacity limit but run with spot-price services. It can be simplified: $\lambda'^{*}_{Hybrid} = arg_{\lambda}\{\int_0^N \tilde{v}x\, f_{\lambda}(x)dx + \int_N^{\infty} \tilde{v}N\, f_{\lambda}(x)dx - P_{Reserved} = 0\}$. Here $\tilde{v}$ is the difference in utilities from a single job execution with reserved and spot-price services.

When there is a capacity limit associated with reserved services contracts, the demand of the marginal client will be higher than that in the case of unlimited reserved capacity. The term $\int_0^N \tilde{v}x\, f_{\lambda}(x)dx + \int_N^{\infty} \tilde{v}N\, f_{\lambda}(x)dx - P_{Reserved}$ is the increase in a client's surplus from using reserved services with limited capacity compared to using

spot-price services. Similarly, the term $\int_0^N \tilde{v}x\, f_\lambda(x)dx + \int_N^\infty \tilde{v}x\, f_\lambda(x)dx - P_{Reserved}$ is the increase in a client's surplus from using reserved services without a capacity limit compared to using spot-price services. The increase in the client's surplus will be smaller when the reserved capacity is limited than when it is unlimited. This applies to all clients. Further, note that $\int_0^N \tilde{v}x\, f_\lambda(x)dx + \int_N^\infty \tilde{v}x\, f_\lambda(x)dx - P_{Reserved}$ is equivalent to $\tilde{v}\lambda^{*'}_{Hybrid} - P_{Reserved}$. This leads to the inequality $\lambda^{*'}_{Hybrid} > P_{Reserved} / \tilde{v} = \lambda^{*}_{Hybrid}$, which suggests that fewer clients will choose to purchase the reserved services contract when capacity is limited.

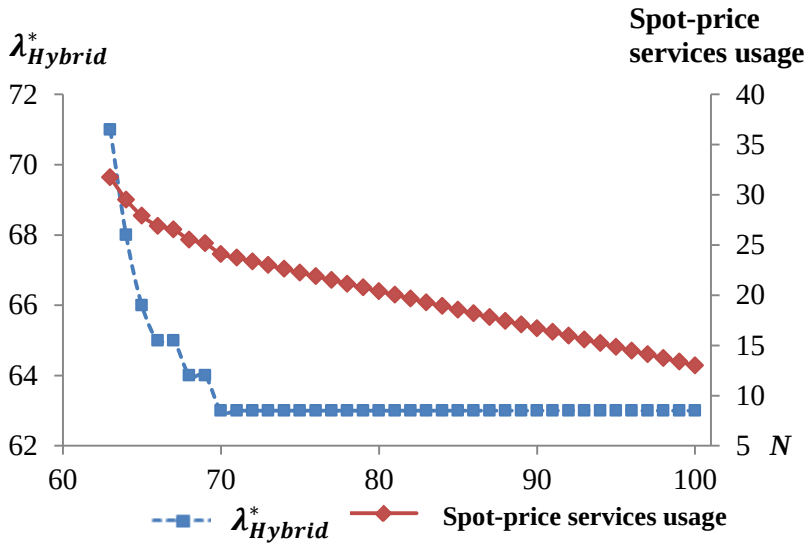
Figure 3.8 shows the market segmentations when the reserved services contract has limited and unlimited capacity. Let λ^{*}_{Hybrid} denote the demand of the marginal client in the case of unlimited capacity, and $\lambda^{*'}_{Hybrid}$ denote the demand of the marginal client in the case of limited capacity, Figure 3.8 shows that $\lambda^{*}_{Hybrid} < \lambda^{*'}_{Hybrid}$. When there is no capacity limit, clients with job arrival rates less than λ^{*}_{Hybrid} will use spot-price services only and the rest will purchase reserved services contracts. When capacity is limited, the client with the job arrival rate λ^{*}_{Hybrid} will no longer find it attractive to buy the reserved services contract. This will also be true for clients with job arrival rates between λ^{*}_{Hybrid} and $\lambda^{*'}_{Hybrid}$. Only those with job arrival rates higher than $\lambda^{*'}_{Hybrid}$ will purchase reserved services contract. They will also use some spot-price services. Clients with job arrival rates lower than $\lambda^{*'}_{Hybrid}$ will use spot-price service only.

Figure 3.8 Market segments



In Figure 3.9 and Figure 3.10, several numerical examples illustrate the impacts of the resource capacity limit on the vendor's profit, usage of spot services, and the market share of reserved services. The parameter values that are used in the numerical examples are shown below the figures. The optimal price of the reserved services contract is $P^*_{Reserved}$ = \$99.42, which is calculated based on the Hybrid Pricing Strategy Proposition (P3). By varying *N* from 60 to 100, it is possible to calculate the corresponding demand of the marginal client in terms of its job arrival rate, the usage of spot-price services, and the vendor's profit. Figure 3.9 indicates that the overall usage of spot-price services decreases with the capacity limit *N*, as does the demand of the marginal client. This suggests that the demand for reserved services will exhibit a monotonic decrease in the capacity limit.

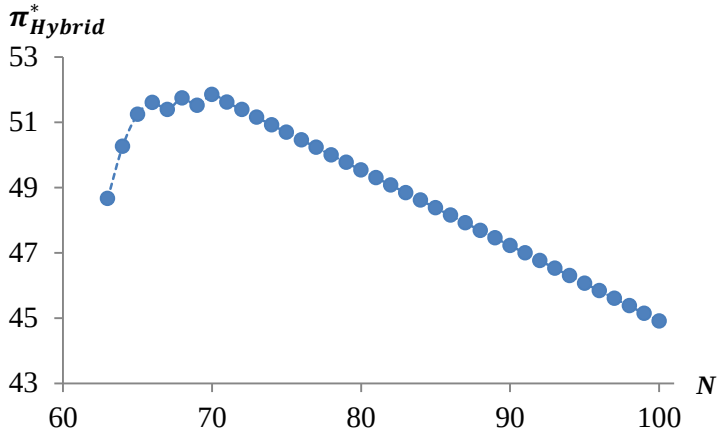
Figure 3.9 Spot-price services usage and the demand of the marginal client



Note: The assumed values in the figure are: $v_L = 2$, $v_H = 4$, $P_L = 0.5$, $P_H = 1$, $Prob_L = 0.5$, $\gamma = 0.3$, and $\Lambda = 100$.

Figure 3.10 shows that the vendor's profit is concave in the capacity limit N. When capacity is limited and above a certain value, the vendor's gain in profit from additional usage of spot-price services will be larger than the reduction in profit due to the decreased sales of the reserved services contracts. Overall profit will start to decrease if the vendor further reduces the capacity of the reserved services contract. As in Figure 3.9, the vendor will start to lose profit when capacity is limited to a value lower than 70.

Figure 3.10 The vendor's total profit when resource capacity is finite in the reserved services contract



Note: The assumed values are: $v_L = 2$, $v_H = 4$, $P_L = 0.5$, $P_H = 1$, $Prob_L = 0.5$, $\gamma = 0.3$, and $\Lambda = 100$.

The next step is to formulate the vendor's optimization problem. The vendor needs to decide the price for reserved services P_{Hybrid}, the capacity limit N, and the probability for a low spot price to occur $Prob_L$, based on:

$$\max_{(P_{Reserved}, N, Prob_L)} \{(1 - F(\lambda^*_{Hybrid}))\, P_{Reserved} + F(\lambda^*_{Hybrid})\, E(\lambda | \lambda < \lambda^*_{Hybrid})\, \bar{P}_{Spot}\}$$

Here $\lambda^*_{Hybrid} = arg_\lambda\{\int_0^N \bar{v}x\, f_\lambda(x)dx + \int_N^\infty \bar{v}N\, f_\lambda(x)dx - P_{Reserved} + \int_N^\infty U_{Spot}(x - N)\, f_\lambda(x)dx = \lambda U_{Spot}\}$ indicates the demand of the marginal client who will be indifferent buying or not buying the reserved services contract. The constraint $Prob_L \geq \gamma v_L / [(1+\gamma)\, v_L - P_L]$ applies because otherwise clients will not submit any jobs to the spot-price services.

Due to the complexity of this problem, it is mathematically intractable. Instead, I derived the solution through a computational approach in the following way. The vendor tests all possible N within a reasonable range, such as (Λ/2, Λ) or

($\Lambda/3$, Λ), and optimizes $P_{Reserved}$ at each value of N. The optimal value of $Prob_L$ can always be found within (0, 1/2), as stated in the Hybrid Pricing Strategy Proposition (P3), or at the boundary 0 or ½.

3.6 Discussion and Implications

Several assumptions in the analytical model are critical in derive the results. In this section, I will first discuss how the results will be affected by these assumptions. I will then provide managerial implications of the results.

The assumption of uniform distribution of job value does not affect the vendor's decision to provide spot-price services. It however may affect clients' job submission behavior. To see this, we need to consider the possibility that the vendor limits resources capacity associated with the reserved services contracts. I assume zero marginal services-provision cost. I then derive Lemma 1 which states that the vendor is best not to limit the resources capacity of the reserved services contract. Without the assumption of zero marginal services-provision cost, however, this assertion no longer holds. The vendor will limit the resources capacity associated with the reserved services contracts. Clients who purchase the reserved services contract will also need to use spot-price services. So, rationale clients will use reserved services to secure as high job value as possible. This is done through a job submission strategy based on a threshold job value v^0: only jobs with value higher than v^0 run with reserved services and the rest run with spot-price services. The threshold job value v^0 is affected by the distribution of job value.

The distribution of job value does not affect vendor's decision on hybrid pricing and its profit, even when different distributions of job value are considered. This is

because of the assumptions of zero marginal service-provision cost and exogenous spot prices. These assumptions relax the vendor from selecting a proper resources capacity for the reserved services contracts, and enable the vendor to only consider clients' expected job value when it prices the reserved services contract. If the model endogenizes spot prices to make the vendor control them, the vendor's decision on the spot prices has to depend on v^0, which represents clients' highest willingness to pay for the jobs run with spot-price services. In this case, whether the vendor can gain higher profit through limiting resources capacity of the reserved services contracts is not definite.

By providing spot-price on-demand services, the vendor gains flexibility in allocating the resources even when clients are using them. This is a big advantage of the vendor that improves the efficiency in resources allocation. This is similar to the common practices in airline industry where airline companies use a centralized reservation system to control the availability of some fare classes based on rules that consider time to departure, seat inventory, origin/destination control, and flight legs, which is in accordance to the best practices suggested by works in revenue yield management (Boyd and Bilegan 2003).

The findings in this research have several managerial implications for cloud computing services vendors. The first implication is related to quality differentiation and price discrimination. Quality differentiation is optimal for vendors if the difference in variable costs to support different quality levels is low (Varian 1997). A vendor may first produce a *superior service*, but later disable some functionality without affecting its usability to offer a lower-priced *inferior service* together with the superior one.

This is the classical *damaged goods strategy*, for example, represented by how Intel sells its computer chips. This strategy is preferred when it costs more to produce different quality levels than to restrict the functionality of the superior services, and will permit clients to self-select what suits them best (Deneckere and McAfee 1996). This strategy seems to be valid for cloud computing services too. Clients will use the superior services for high-value jobs and the inferior services for low-value jobs. The vendor will need to be careful though: pricing the superior services too high or the inferior services too low will reduce their profitability.

Spot-price on-demand services are subject to services interruption risk. This will diminish the valuation that clients put on such services. They are damaged services in the IT services context. I call the provision of spot-price services that are subject to services interruption *damaged services strategy*. The probability of services interruption is essentially a quality differentiator. The findings of this study suggest that the vendor should not minimize the probability of services interruptions. This indicates that, even when the vendor has extra resources and it incurs zero cost to serve additional clients, it is better off to hold back some resources from spot-price services. It is an approach the vendor can use to control the services interruption risk, which affect clients' expectation of quality of the spot-price services. This somewhat paradoxically runs against the grain of the customer service focus in services management. We have seen that when clients are sensitive to services interruption, achieving low interruption levels may not be worthwhile. With a non-monopoly market structure, a competitor's pressure is likely to make it necessary for the cloud computing services vendor to configure a low level of services interruption risk. A cloud computing ser-

vices vendor can learn from the findings of this research, but will still need to be cautious since few are monopolists with respect to the spectrum of their services offering. They may not be monopolists for very long in a market that rewards IT services innovation.

A third implication is that the hybrid pricing strategy with services interruption as a quality differentiator may benefit vendors only when their clients are sensitive to services interruptions, or spot prices are relatively high compared to the value of the clients' jobs. These two conditions seem contradictory because clients will not be willing to pay a high price for spot-price services when they are highly sensitive to services interruption. In this circumstance, they will find fixed-price reserved services more attractive compared to spot-price on-demand services. This will trivialize the cannibalization effect caused by introducing spot-price services to the market. The nature of computing jobs that run with cloud computing services will affect clients' risk sensitivity to services interruption.

For example, businesses that are built on cloud computing services will require some level of continuity in the computing resources. They will encounter serious operational problems if their cloud services instances are interrupted. Clients who conduct scientific computing jobs, however, will care more about their budget and probably have flexible completion time requirements. They will be less sensitive to services interruptions than business clients. So the hybrid pricing strategy may work well where a majority of clients have similar preferences for the continuity of their services. Otherwise, it will not necessarily result in greater profitability for the vendor.

The second and third implication show how and when the cloud services vendor

benefits from providing the spot-price interruptible services. To manage the interruption risk is important. The services interruption risk should generally increase with clients' sensitivity to the interruptions. Based on the numerical investigations presented in Section 3.4, a rule of thumb is to configure the services interruption risk between 0.2 and 0.25. I derive rule, however, based on the assumption of clients being homogeneous in their sensitivity to services interruptions. If we consider a more realistic scenario where clients are heterogeneous in this aspect, only the clients with high demand and high sensitivity to services interruptions will purchase the reserved services contract, and these with low sensitivity to services interruptions, no matter their demand are high or low, will use spot-price services. In this case, the market segmentation changes, and fewer clients will purchase the reserved services contract. Therefore, the vendor should configure an even higher services interruption risk.

I have discussed the managerial implications of the services interruption risk associated with spot-price services, treating it as a quality differentiator. It is critical in the damaged services strategy. The services interruption, however, is more than a quality differentiator. It also enables the cloud services vendor to control its inventory of compute resources to gain higher profit. For example airline industry has practiced seats inventory control: they will open some fare classes at certain time and close them after a time period controlled by a central reservation system. (Williamson 1992, Boyd and Bilegan 2003). For example budget airlines often offer zero priced tickets for travels during off-peak seasons, and provide tickets with big discount for multi-journey flights during peak seasons. Similar practice can be found in hotels where hotel managers control room inventory to provide different room-type and price-

category during peak and off-peak seasons to maximize revenue (Relihan III 1989).

Following the first two implications, the findings of this research also have implications for how a cloud service vendor can use capacity, together with the appropriate pricing strategy, as a tool to improve profit. I have discussed this and used a numerical example to illustrate the efficacy of a capacity limit for the reserved-services contract in improving the vendor's profit. The key is to let the vendor ration its high-quality services to clients with high demand and to let these clients use their limited reserved capacity for high-value jobs and use spot-price services for the rest of their jobs. Vendors will get more surplus from clients with high demand, and such profit gains will be at the expense of consumer surplus and overall social welfare. This is because more jobs will be run with spot-price services, facing the risk of services interruption and leading to loss in consumer surplus and social welfare.

A final implication relates to how to manage service interruption risk from the vendor's perspective. It is critical to the effectiveness of the hybrid pricing strategy that clients perceive quality difference between the fixed-price reserved services and the spot-price on-demand services. Only when clients perceive the quality difference between these two types of services, the vendor can charge the reserved services a high price, so that it can reduce the cannibalization effect by the spot-price services yet maintain a high profit from the fixed-price reserved services. In practice, clients are unaware of the potential losses associated with interruptions. They also may not know how to estimate their likelihood. Although Amazon makes available historical spot prices for EC2 spot-price instances for the past 90 days, it may not be enough information for the client to manage its risk effectively. So many clients are likely to

over-estimate the risk of services interruption, and will perceive a low value for spot-price services. As a result, the vendor will likely enjoy less profit. To ameliorate such impacts on profitability, a forward-looking vendor may wish to act as a *risk management infomediary* for IT services with respect to its outcomes. Or an outsider can take this role, just like the travel agents in the airline industry and hospitality industry that facilitate the booking of air flight tickets and hotel rooms. Similar uncertainty issue arises when a client uses spot-price services. I will discuss related issues, including the characteristics of the uncertainty associated with the use of spot-price services, whether a client can gain sufficient information to eliminate the uncertainty, and how will the uncertainty affect clients' valuation of the spot-price services, in the next chapter.

3.7 Conclusion

This research analyzes a monopolist's pricing strategy for cloud computing services. Using hybrid pricing, the vendor takes services interruption as a quality differentiator to offer two types of services to the market: fixed-price reserved services and spot-price on-demand services. The latter is associated with a certain level of services interruption risk and will be viewed as lower-value services. The use of hybrid pricing in the cloud computing services market is novel. To the best of my knowledge, this is the first study to analyze its economic impact and to offer practical pricing strategy recommendations to cloud vendors. This research shows that, only when (1) clients are highly sensitive to service interruptions or (2) the expected spot price is high relative to the value of clients' jobs, should the vendor employ the hybrid pricing strategy. It will lead to lower consumer surplus and lower social welfare in most cas-

es. On the other hand, the results also show that the vendor is able to increase its profit by keeping the services interruption risk at a high level, so that the cannibalization effect between fixed-price reserved services and spot-price on-demand services will be minimized.

This work is subject to several limitations. In the current model, spot prices for on-demand services are assumed to only have two exogenous values (high and low), and the vendor decides the probability of each value to occur. This, in turn, determines the level of service interruption. In business, spot prices will be dynamically determined by real-time supply and demand. The cloud services vendor will be able to collect its clients' price bids and set spot prices appropriately. This decision-making problem for the vendor is not analyzed here. Moreover, clients cannot delay the submission of jobs.

I also do not consider whether two spot prices are incentive-compatible. When the client can delay job submissions, it will evaluate a job runs with spot-price services at different price levels, and submit the job if the current spot-price yields the highest net utility, or wait until the spot-price level changes. An incentive-compatible pair of spot prices will induce the client to separate job arrivals according to their value, and submit high-value jobs when spot prices are at a high level and low-value jobs when spot prices are at a low level. This will change the client's overall expected utility from spot-price services, which will affect its willingness-to-pay for the reserved services contract as well.

In addition, the arrival of jobs is assumed to be independent and identical over all of the periods. In reality, however, job arrivals are not the same across different peri-

ods. Clients will experience peak and non-peak job-arrival periods, and therefore their decisions will be more complicated. Despite these simplifying assumptions, my model is able to capture the key features of hybrid pricing strategy in the cloud computing services market, and offers meaningful recommendations to the vendor.

Furthermore, I implicitly assume that all clients are well informed about the services interruption risk. Therefore they are able to perceive the quality difference between the fixed-price and spot-price services. This is not always true in reality. Clients may over- or under-estimate the services interruption risk associated with the spot-price services. This may lead to ineffectiveness of the hybrid pricing strategy. To avoid this, the vendor should help clients perceive the quality difference between fixed-price and spot-price services. For example, the vendor can provide risk-analysis support to clients to eliminate biases in their assessment of the services interruption risk.

The limitations suggest directions to extend the analytical model towards a more general and realistic setting. For example, it is possible to relax the assumption of exogenous spot prices. Another possibility is to allow the job-arrival rate to vary over time. By studying the interactions between IT services vendors and clients, I expect to understand the cloud computing services market more fully and make more meaningful managerial recommendations.

4 An Experimental Study of a Compound Pricing Mechanism for Brokered Cloud Services

The prior chapter discussed a cloud services vendor's decision about offering fixed-price reserved services or spot-price on-demand cloud computing services. The analysis was based on a model assuming that all clients are aware of the risk of services interruption associated with the spot-price services. They also can calculate the level of this risk. The simplifying assumption in the analytic model that spot prices take only two values, which is necessary to get insights on the problem setting while not making the model too complicated to analyze, is not too unrealistic.

In reality, taking Amazon EC2 spot prices as an example, there are multiple price levels and predicting the spot price that will occur 5 minutes later is a burden for clients. In fact, clients can hardly predict whether and when their jobs run, when spot-price services will be interrupted. This will leave the clients in an uncomfortable situation, but a new opportunity for business is available. Services that can help clients eliminate the uncertainty associated with the spot-price services will create value for clients. How clients value such services and what factors will affect their valuation, however, are worth investigating before the real business initiatives taking place. This chapter will present an experimental study that addresses these issues.

4.1 Introduction

Different approaches to the pricing and sale of cloud computing services have evolved over the years. Vendors have sold computing and storage services with fixed prices when their clients need them. This has been called the *on-demand* model. I will refer to it as the *fixed-price model* to emphasize the role that prices play in cloud

computing. In December 2009, Amazon.com launched demand-driven dynamic pricing as an alternative (Higginbotham 2009). This approach has come be known as the *spot-price model*, and changes in market prices are reflected for jobs running as *spot-price instances*. Under this mechanism, a client's access to a specific instance of a computing resource depends on its *bid price*, the maximum price per unit of services that the client is willing to pay. Changing vendor services supply and clients demand jointly decide spot prices in the market over time.

Even though the spot-price model has been attractive to clients that wish to avoid over-paying for services under the fixed-price model, my ongoing observations in a field study of cloud computing services during the past several years suggest that spot prices can be volatile. One example is the market's experience with Amazon.com's spot prices during 2011, on the basis of one type of instance – the m2.2xlarge instance from the eastern region of the United States. As reported by the Cloud Exchange (defunct) (Brandon 2011), the price of this instance type rose precipitously to US$999.99 for a two-hour period on September 26, 2011 from US$0.44 per hour on average. This caused most jobs that were being processed at that time to be interrupted, since few clients previously indicated such a high reservation price.

Although such prices are not typical, a number of questions arise: (1) How important is it to avoid the interruption of services? (2) How much should a client be willing to pay to obtain a services guarantee? (3) How will client willingness-to-pay be affected when the market demand leads to a relative shortage of supply of cloud computing services and escalates prices in the short term, but the historical trends in longer term are known? In particular, clients differ in their abilities to interpret and

predict the short-term price fluctuations. This will cause clients to assess the business value of their jobs and to identify an appropriate bid price for the services they need, as services with a job completion guarantee, when market demand and spot prices change.

Digital intermediaries provide well-known capabilities to mitigate the risks that arise around the exchange of products and services with buyers and sellers in market settings (Malone et al. 1987). This is true in financial markets for stocks, bonds, commodities and foreign exchange, and in the markets for TV and radio broadcasting bandwidth, among other settings. Similar opportunities exist in the market for IT services, where there is a need for price discovery, vendor and client matching, the creation of stakeholder informedness, the diminution of related transactions and operations costs, and risks of exchange.

Cloud computing services brokers are natural intermediaries to mitigate the risks of exchange and support improvements in the quality of markets for IT services (Jackson 2012). They can aggregate resources to lower the risk of a client's sudden loss of services, and customize services to address different client needs (Gartner 2011). This is similar to how wealth management services brokers support the varied investment needs of high net worth clients – a term commonly used in the financial sector, referring to those having high investable financial assets excluding their primary residence. This includes, for example, clients with financial assets in excess of US$1 million. Multiple financial institutions may have limited services or services delivery capabilities to address the full spectrum of client needs. In addition, research works have been conducted to improve the efficiency of resources allocation in virtu-

alized computing environment, including consolidation of virtual instances, better planning and more efficient management of capacity of the virtualized data centers (Speitkamp and Bichler 2010, Aedagna et al. 2013, Ghosh et al. 2013, Kesavan et al. 2013). Some brokered services vendors may wish to design, build and offer their own complementary services to further mitigate the negative effects of services interruption. For example, PiCloud (www.picloud.com), a cloud services broker, leverages Amazon.com's spot services to implement a pool of job queues to deliver services faster than typical infrastructure services while reducing costs for its clients. It requires clients to use a specific programming language to interact with the cloud platform. Then, in case of services interruption, the vendor's cloud platform can take over the interrupted job and restart it autonomously.

The major concerns these cloud services brokers address are related to the continuous availability of cloud services and management of computing resources from several different cloud services vendors. There have been very limited attempts, however, to investigate the differences among clients that affect their economic behavior. This will allow efficient cloud computing brokerage services and resources allocation mechanisms to be designed for them.

This research proposes a compound pricing mechanism that is flexible and emphasizes clients' concerns about the risks associated with cloud computing services. In the proposed mechanism, the vendor, as a risk-mitigating digital intermediary, offers spot-price services that are configurable by the client to address its resource requirements, deadlines and performance constraints, the price it will pay, and the compensation it will receive if the vendor fails to perform. The vendor, by the same token,

will have the desired flexibility to evaluate and accept or reject bids from clients. This is a risk-based value exchange between the vendor and its client, performed in a way that indemnifies the client against financial losses that may arise when job completion deadlines are not met due to services interruption.

For the proposed mechanism design to be viable, assessing clients' willingness-to-pay for services based on the related job duration risk, client risk informedness and client risk characteristics is necessary, along with the different prices they must pay. But how will these factors affect a client's willingness-to-pay for cloud computing services that include a job completion guarantee? This study uses a web-based experimental testbed to carry out an empirical study to supply the answers.

The results in this study show that client risk informedness and client risk aversion affect their willingness-to-pay for the services that are offered through the experimental testbed. Understanding the interaction between client risk informedness and client risk aversion is useful for the brokered services vendor. This is especially true for clients with a higher level of risk aversion: when they have a better knowledge of the job completion risk, they will exhibit a lower level of willingness-to-pay for brokered cloud computing services that are supplied using the compound pricing mechanism. Clients who are less risk-averse are less likely to value services with guaranteed job completion. These results suggest that the cloud vendor or broker should be strategic in affecting client risk informedness. This research contributes to the literature on cloud computing by proposing a novel mechanism design for brokered cloud computing services using the compound pricing mechanism.

4.2 Literature and Development of Hypotheses

The brokered cloud services that this research proposes are additional offering to the fixed-price and spot-price cloud services already available in the market. Fixed-price reserved services set aside cloud computing resources for clients, and guarantee job completion. Spot-price services do not provide such job-completion guarantees. There will be a loss of efficiency because clients who need a moderate level of job-completion guarantee and have moderate capabilities to deal with the lack of a guarantee may stay out of the market. To gauge the economic potential of brokered cloud services, it is important to know how much clients are willing to pay for the brokered services and how client willingness-to-pay will be different in the presence of several key variables.

Many methods have been developed to measure client willingness-to-pay, either directly or indirectly, in marketing literature. See Miller et al. (2011) for a comprehensive review. This study uses a direct approach. The focus of this research, however, is not on how to measure client willingness-to-pay; I am more interested in the factors that potentially will affect client willingness-to-pay, and are unique and prominent in the context of cloud computing. One such unique factor is services interruption-related issues created when spot-price services are provided.

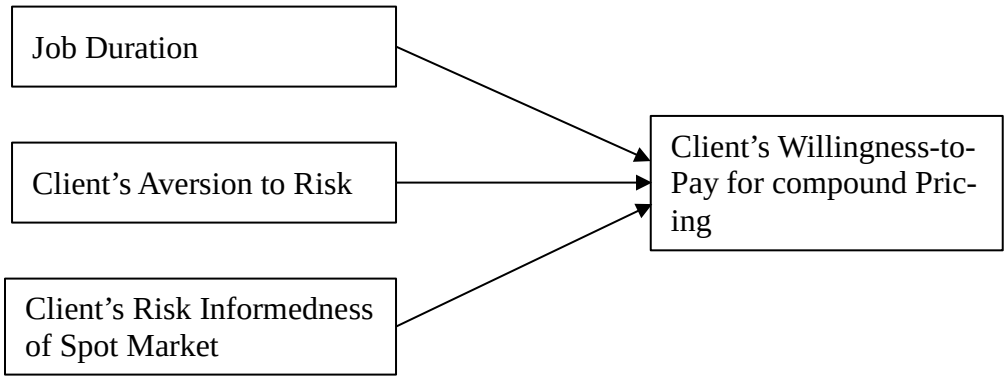
A client's willingness-to-pay for a product is essentially its estimate of the value of that product. When evaluating the value of the product, clients tend to have a bias toward the estimate that is initially given to them (Tversky 1974). In the context of cloud computing, fixed and spot prices act as *reference points*. To estimate a client's willingness-to-pay involves a comparison between these prices. The subsequent esti-

mates will be biased toward the price that is preferred by clients. Clients seek the lowest prices in the market; so, spot prices will be the main reference point, on which a client can place *decision weights* to represent its willingness-to-pay (Kahneman and Tversky 1979), and the fixed price will be the upper bound of the client's willingness-to-pay. In addition, because spot-price services are subject to unexpected services interruption, a client's willingness-to-pay can be viewed in terms of *certainty equivalents* relative to the financial losses that may arise when services become unavailable to the client.

Expected utility theory, discussed by Fishburn (1988), suggests that the certainty equivalent of an asset with an uncertain outcome should be calculated by a weighted sum of all possible outcomes, based on their probability of occurrence. There is much controversial evidence regarding the axioms of utility theory, however. As a result, alternative theories, such as prospect theory and subjective expected utility theory—which suggest the weights to be associated with potential outcomes should be based on, but different from, their probabilities—have been proposed to explain clients' decision-making amid uncertainty (Kahneman and Tversky 1979). Therefore, to understand a client's willingness-to-pay for brokered cloud services, we need to know (1) the probability of spot-price services interruption, and (2) how a client adjusts its weights in decision-making based on the probabilities.

I will focus on variables that affect a client's perception of risk and willingness-to-pay (see Figure 4.1), and offer several related hypotheses.

Figure 4.1 Research model



Job duration and completion risk. A client's willingness-to-pay for cloud computing services will vary based on the purposes to which they are put: for example, running a two-hour computing-intensive scientific simulation, or hosting an e-commerce website will be valued differently.

To get a sense of the empirical regularities of spot-price cloud computing services, I obtained time-series data on prices from the Amazon Elastic Computing Cloud for the period from January 31, 2012 to January 31, 2013. (See the Appendix C1.) I found that the number of services interruptions was positively associated with job completion duration. In the literature on risk and uncertainty, Krupnick et al. (2002) found that clients were willing to pay to reduce their risk of mortality, and their willingness-to-pay increased with the magnitude of the likely reduction. People also purchase insurance products to hedge against unexpected financial losses. Similarly, experimental jobs with a higher interruption risk should induce clients to pay more in order to eliminate the risk. In addition, the compound pricing mechanism provides guaranteed job completion at a price, including a compensation scheme, and this will ensure a client against financial losses due to unexpected services interruptions. This should influence the client's willingness-to-pay for such services when the interrup-

tion risk it faces is high:

- **Hypothesis 1 (Job Duration and Completion Risk):** *Client willingness-to-pay for cloud computing services with a job-completion guarantee provided under the proposed compound pricing mechanism will be higher for jobs of longer duration that are more subject to job-completion risk.*

Economic theory supports this hypothesis. Utility theory suggests that a client will pay a price premium for protection against risk. Risk in the cloud computing services setting arises due to the possibility of services interruption when jobs run with spot-price services. Decision-makers often weigh outcomes differently from their probabilities. In addition, decision-makers often over-estimate or under-estimate the probabilities associated with risky outcomes (Kahneman and Tversky 1979), and this bias typically is consistent across different kinds of risks (Schoemaker 1993).

Client risk aversion. Decision-makers typically prefer certain outcomes to ambiguous ones (Schmidt et al. 2008). Cloud computing services that involve spot-price services with guaranteed job completion generally will be preferred to spot-price services without guarantees. This is due to client risk aversion. As a result, the client will be willing to pay a higher price for the guaranteed completion of a job, subject to its risk aversion for financial losses. This suggests:

- **Hypothesis 2 (Client Risk Aversion):** *Clients who are more risk-averse will have a higher willingness-to-pay for cloud computing services with a job-completion guarantee provided under the proposed compound pricing mechanism.*

A decision-maker's sensitivity to risk is influenced by situational factors, such as

the manner of presentations of different possible outcomes (Slovic 1972b). The situational factors affect the cognitive effort that a decision-maker has to put into assessing the risk. Therefore, to examine the effect of a decision-maker's risk aversion on its decision-making, the influence of situational effects should be controlled. Prior experimental research has explored the relationship between different bidder types defined by their aversion to risk and their bidding behavior in a controlled laboratory experiment (Bapna et al. 2010). More specifically, decision-makers who are risk-averse tend to favor certain outcomes more than those who are risk-seeking (Schoemaker 1993). On the other hand, clients as decision-makers tend to over-estimate the probability of losses even though the actual probabilities are small (Kahneman and Tversky 1979). The bias of estimating uncertain outcomes with small probabilities tends to be moderated by the decision-maker's attitude toward risk. Decision-makers who are more risk-averse perceive a higher probability of loss than those who are less risk-averse, although both estimates are likely to be higher than the actual probability. Therefore, no matter whether expected utility theory or prospect theory describes a client's utility more appropriately, the effect of a client's risk aversion on its willingness-to-pay should be consistent.

Client risk informedness. When consumers are not fully informed about the quality or performance of a product or service, they apply an uncertainty discount (Li et al. 2013), which diminishes their willingness-to-pay (Clemons and Gao 2008, Markopoulos and Clemons 2013). This applies to client informedness and willingness-to-pay in cloud computing services too. A greater level of informedness will lower a client's uncertainty discount for the value of the services. Historical information on spot

prices may influence a client's perception of the need for purchasing services that include guaranteed job completion. Sharing critical information that shows the client the job-completion risk, based on the increased likelihood of jobs with longer durations not reaching completion, will help to inform the client, for example. This leads to:

- **Hypothesis 3 (Client Risk Informedness):** *Clients with a lower level of risk informedness about their jobs will have a higher level of willingness-to-pay for cloud computing services with a job-completion guarantee provided under the proposed compound pricing mechanism.*

In the context of this research, the probabilistic services interruption is the major source of client uncertainty about spot-price services. Risk informedness determines to what extent a client's assessment of the certainty equivalent of the financial losses due to services interruption will be accurate. Bear in mind that spot prices usually fluctuate around a base price, and services interruption only happen when spot prices rise to higher levels. So clients with limited risk informedness will likely get estimates that deviate more from the base spot price than clients who are more informed. So clients with limited risk informedness should be willing to pay a higher price for protection against services interruption.

4.3 The Experiment

A vast body of literature in economics suggests that clients may be uncertain about product performance, and also their preference structures (Urbany et al. 1989, Gregory et al. 1993, Wang et al. 2007). In this experimental research, client uncertainty about preference structure is controlled for by giving explicit instructions on

the goals and performance measures in the experiment so that price changes for spot-price services and the completion status of jobs run with spot-price services are the only sources of uncertainty. Real-world spot-price data represent price changes for the hypothetical spot-price services in the experiment. It also involves the manipulation of the level of uncertainty associated with spot-price services by providing subjects with different information about the uncertain aspects of the services.

Two pilot sessions occurred before the formal experiments. The first was done in November 2012. Six full-time employees in a Singapore-based high-tech research institution participated. Based on the feedbacks, I simplified the testbed's interface and then conducted a second pilot session in December 2012. Eight post-graduate students from IT-related majors participated. They commented that the calculations of the costs and benefits were complicated and difficult. This motivated me to further refine the design of the experimental jobs and improve their descriptions.

After the two rounds of pilot sessions and a series of refinements of the testbed interface, experiment instructions, and experimental job descriptions, I conducted the experimental sessions in February 2013. Working professionals from a Singapore-based public research institution that focuses on science and technology research formed the subject pool. The participants were knowledgeable about business analytics and IT services, a requirement for their involvement. Participants received S$20 for joining the experiment, plus a performance-based bonus that ranged from S$10 to S$50 to induce rational economic decision-making (Smith 1976). During the study period, S$1 was about US$0.80. 45 subjects completed the experimental procedure. They had a mean age of 34 years, and 71% were male. Their levels of experience

confirmed that no participants had truly expert-level knowledge, but nor did they lack domain knowledge or managerial decision-making experience. (See Figure 4.2.)

Figure 4.2 Experimental design

		Job duration		
		Three hours	**Five hours**	**Ten hours**
Risk informedness	**Low**	21 subjects		
	High	24 subjects		

Note: *Job duration* is a within-subjects design variable. Subjects all were assigned the same set of three computing jobs with a duration of 3, 5, and 10 hours.

Experimental set-up

Services interruption risk was increasing in the duration of the jobs that ran with spot-price services. I manipulated *job duration* was configured as a within-subjects variable. Different job durations induced different level of services interruption risk. Subjects all were assigned the same set of three computing jobs with three, five, and ten hours duration. A subject's *risk aversion* was measured after all the experimental sessions were completed, as way of minimizing the subject expectancy effect, using a questionnaire adapted from Weber et al. (2002). (See Appendix C2.) *Risk informedness* was manipulated using two randomized between-subject conditions. In the *low risk-informedness condition*, subjects were provided with historical spot services price information, which matches the kind of information supplied in the spot-price, on-demand cloud computing services market. In the *high risk-informedness condition*, subjects were given additional risk analysis support, beyond historical spot price information, to help them evaluate the costs and benefits of different cloud computing services.

Implementation of the experimental testbed

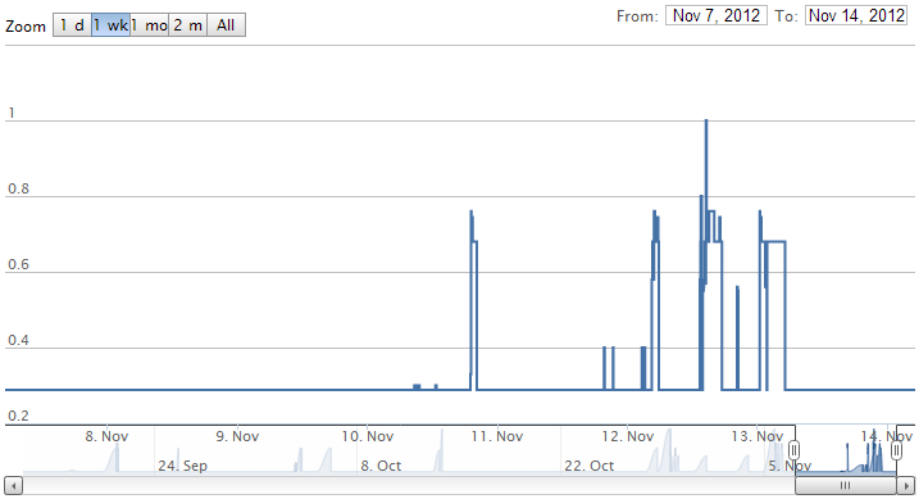
This research involved the design of a website called SmarterCloud (www.smarter-cloud.biz) to represent a cloud computing services vendor and its compound pricing mechanism for the services offerings. The testbed was implemented with Python on the Google App Engine platform-as-a-service. (See Figure 4.3.) The site allows clients to compare and purchase cloud computing services in the following way. For the purchase of fixed-price services, all subjects had access to job cost information. For spot-price services, statistics on hypothetical purchases were derived from a simulation of experimental jobs with real historical spot prices. Subjects received them based on their assignment to different treatment groups. For a description of the simulation, see the Appendix C1. For cloud computing services with guaranteed job completion, the vendor offers the client compensation if there is a problem, but the client must pay for the guarantee. This involves a simple best-offer algorithm with a termination rule over two rounds at the most in which the participant makes a bid for the job. The experimental vendor is programmed to reject the first bid, and then will use the bid in second round to assess a subject's willingness-to-pay.

I treat *job duration* as a within-subjects variable. Subjects were all assigned the same set of three computing jobs of a duration of 3, 5, and 10 hours. A subject's *risk aversion* was measured after all the experimental sessions were completed, as a way of minimizing the subject expectancy effect, using a questionnaire following Weber et al. (2002). *Risk informedness* was operationalized using two randomized between-subject conditions. In the *low risk-informedness condition*, subjects were provided with historical price information on spot-price services, which matched the infor-

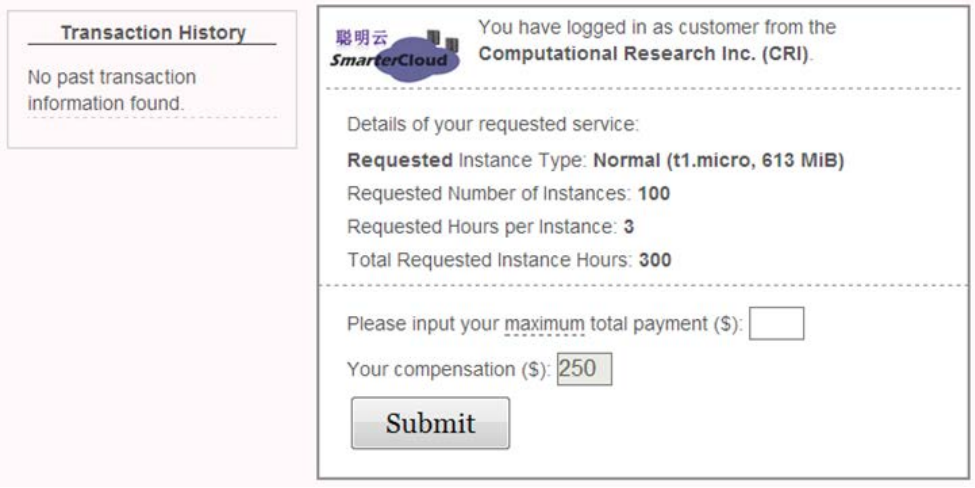
mation supplied in the spot-price, on-demand cloud computing services market. In the *high risk-informedness treatment condition*, subjects received additional risk analysis support, in addition to the historical spot-price information, to help them evaluate the costs and benefits of different cloud computing services. (See Table 4.1.)

Figure 4.3 Experimental testbed: SmarterCloud (www.smarter-cloud.biz)

a. Historical spot price shown to participants without risk analysis support



b. Transaction interface through which subjects make offers to SmarterCloud



Experimental jobs, measurement of job performance and willingness-to-pay

Subjects were told to run a series of simulation jobs with cloud computing services of three different durations: 3, 5, and 10 hours. This parameter setting conforms to typical jobs in the real world data centers, where 99% of scientific computing tasks have durations of less than ten hours (Iosup et al. 2008). The deadline for completing the simulation results was 24 hours.

To ensure that subjects established a perceived cost for spot-price cloud computing services, their performance was evaluated according to the following rules. If a subject made a successful offer to use the services, his or her performance would be calculated based on the amount of profit generated from the jobs, less the payment required for guaranteed job completion. If a request for the services were unsuccessful, then the subjects would have to use spot-price services to execute the jobs.

This was operationalized by randomly generating jobs ran with spot-price services and using their associated completion and revenue outcomes. The subject's performance on each job was calculated based on the amount of profit generated from the execution, less the payment required for spot-price, on-demand services. When a job running with spot-price services was completed, the firm's revenue was S$1,000. When the job was not completed in time, the revenue was prorated based on the completed percentage of the interrupted job, plus a S$200 penalty.

Table 4.1 Risk analysis support for the high risk-informedness condition

	Job completed?	Occurrence	Total payment	Expected profit
On-demand	Yes	100.0%	S$240.00 (S$0.80 × 300)	S$760.00
	No	0.0%	--	--
Spot price-based historical purchases	Yes	98.9%	S$87.74	S$912.26
	No	1.1%	--	S$184.44

Notes: Historical purchase information was available to subjects in the baseline and treatment conditions; similar information for spot-price services was available only to subjects in the related treatment condition.

Experimental procedure

When subjects came to participate in the experiment, they logged onto the SmarterCloud website and were randomly assigned to one of the two risk-informedness conditions: high or low. Each subject read and signed an informed consent form, read the instructions, and took a quiz to test if he or she had a good understanding of the purpose and procedures of the experiment. Each subject then purchased cloud computing services to run three jobs. For each job, a subject was required to initially submit the maximum price that he or she was willing to pay for the services, and then to update the price in a second round of bidding. After the experimental session, the subjects were asked to complete a questionnaire to provide feedback on their experience with the experimental testbed. Another questionnaire was distributed one week later to assess each participant's risk aversion. (See Appendix C2.) The delay in the distribution of the second questionnaire was to minimize the subject expectancy effect.

I conduct a number of control checks. The results suggest that the subjects' ages, experience levels and risk aversion did not differ between the baseline and treatment

conditions. Nor did I obtain evidence through regression analysis that risk aversion differed as a function of presenting or withholding risk analysis support to subjects. This allowed me to examine risk aversion as a stable predictor of a subject's propensity to take risks.

4.4 Data Analysis and Results

Control Check

The control variables need to be assessed to find out whether participants were randomly assigned to one of the two experiment groups. Table 4.2 provides descriptive statistics on participants' demographic information. The results of the check of the control variables suggest that participants' risk propensity, age, and experiences did not differ between the baseline group and the treatment group, confirming that the randomization is effective.

Table 4.2 Characteristics of subjects (*N* = 54)

	Mean (Std. Deviation)			
	Control Group	**Treatment Group**	**All**	***p*-Value (Between-Group)**
Risk propensity	-0.57 (8.10)	0.62 (9.10)	0.07 (8.56)	0.645
Age	33.86 (6.18)	34.88 (6.20)	34.40 (6.14)	0.585
Working experience	2.52 (1.17)	2.96 (1.33)	2.76 (1.26)	0.255
Decision-making experience	1.33 (0.66)	1.54 (0.78)	1.44 (0.73)	0.342
Cloud services usage	1.76 (0.77)	1.83 (0.92)	1.80 (0.84)	0.780
Negotiation experience	1.76 (0.77)	2.25 (1.11)	2.02 (0.99)	0.099
Analytics experience	1.19 (0.40)	1.33 (0.76)	1.27 (0.62)	0.445

Hypotheses Testing

To test the Job Duration and Completion Risk Hypothesis (H1), I conducted a repeated measures analysis of variance. There is no significant effect of job duration on client willingness-to-pay ($F(1.71, 73.47) = 1.53$, $p = 0.23$).[1] For robustness check, I also tested the effect of job duration on client willingness-to-pay in sub-groups, and obtained similar results for both the baseline group ($F(1.56, 31.21) = 0.664$, $p = 0.486$) and treatment group ($F(2, 46) = 2.85$, $p = 0.068$). So H1 is not supported. A plausible reason is that the range of job durations in the experimental setup was insufficient for subjects to consider them as being different.

To explore this possibility, it made sense to conduct a *post hoc* analysis to examine the impact of job duration in cost terms. In this analysis, the unit cost for spot-price services for the three experimental jobs was computed. The aim was to find out whether there were perceivable differences between the unit costs. Based on the risk analysis information in the experiment, the unit cost for spot-price services of a single experimental job was calculated by dividing the expected total cost, including both successful and unsuccessful cases, by the number of instances required and job duration in hours. A client's unit costs for spot-price services are S$0.315, S$0.310 and S$0.306 for job durations of 3, 5 and 10 hours. Subjects probably ignored the differences, since they were not sufficiently large to be meaningful.

In addition, historical spot prices for cloud computing services are helpful for giving a sense of the frequency of their changes within a time period: for instance, 20

[1] Mauchly's test indicated that the assumption of sphericity was violated for the data ($\chi^2(2) = 11.41$, $p = 0.003$). I corrected the degrees of freedom using Huynh-Feldt's estimates of sphericity ($\varepsilon = 0.85$) as a basis for these results.

price changes over 3 months or 100 prices changes in a year. This type of information is not helpful in assessing the interruption risk associated with a specific computing job though. Subjects may not be able to differentiate the risk levels associated with the three experimental jobs. They then may ignore the difference in job durations and not adjust their willingness-to-pay.

I next assessed the Client Risk Aversion (H2) and the Client Risk Informedness (H3) Hypotheses. I conducted an analysis of variance (ANOVA) test, which suggests that risk informedness has a significant effect on willingness-to-pay for cloud computing services under the compound pricing mechanism ($F = 7.41$, $p = 0.009$). In addition, this effect was more pronounced in the high job-risk case ($F = 14.48$, $p < 0.001$) than in the low job-risk case ($F = 4.44$, $p = 0.041$).

I further tested whether the effect of risk informedness is consistent in the high risk aversion and the low risk aversion groups. *High risk aversion* is defined in terms of a subject's risk score being less than -4, which is the mean of the overall risk score minus one half of one standard deviation. Similarly, *low risk aversion* is defined in terms of a subject's risk score being more than 4, which is the mean of the overall risk score plus one half of one standard deviation. Interestingly, the effect of risk informedness is found significant in the high risk aversion group ($F = 11.30$, $p = 0.001$), but not in the low risk aversion group ($F = 0.63$, $p = 0.432$).

The results suggest that the effect of risk informedness on client willingness-to-pay is consistent and robust across different levels of job risk, but inconsistent among subjects with different levels of risk aversion. Subjects with low risk aversion are indifferent in their willingness-to-pay – with or without risk analysis support. The effect

of risk informedness may be dominated by the subjects' inclination to pursue a higher profit, even though there is uncertainty. Alternatively, they simply may have ignored the additional information provided by the risk analysis support tool.

I ran a regression to assess the marginal effects of risk aversion and risk informedness, and found that the main effects and interaction effects were significant. (See Table 4.3.) Sub-sample analysis for cloud computing services jobs with low risk (durations of 5 hours) and high risk level (durations of 10 hours) showed that the main and interaction effects were significant. (See Table 4.3.) Interestingly, the coefficient estimates for the effect of risk informedness suggest a greater impact when job risk is high than when job risk is low. These results indicate that subjects with high risk informedness may perceive larger differences in the level of risk associated with jobs of different durations than subjects with low risk informedness. The repeated measures test results further confirmed that there was no significant difference between client willingness-to-pay under different job durations. Hypotheses 2 and 3 thus are supported.

Table 4.3 Basic model with full sample, low and high job-risk groups

Model variables	Full sample	Low job risk	High job risk
Risk informedness	-0.305*** (0.025)	-0.290** (0.042)	-0.483*** (0.041)
Risk aversion	-0.453*** (0.002)	-0.507** (0.004)	-0.516** (0.004)
Risk informedness × Risk aversion	0.301** (0.003)	0.419* (0.005)	0.357* (0.005)
Adj.-R^2	17.1%	14.0%	31.5%
Note: Standard deviations are reported underneath the coefficient estimates. Coefficients are standardized. *** for $p < 0.001$, ** for $p < 0.05$, * for $p < 0.1$.			

In order to remove the effect of subjects' prior experience with (1) cloud computing services, (2) decisions made under uncertainty, and (3) analytical tasks, I constructed and tested an extended regression model. The extended model includes measures of subjects' experience with working, business decision-making, business analytic tasks, business negotiation activities, and cloud computing services usage. The results are reported in Table 4.4. The interaction effect was weaker in the extended model. In addition, the results for the extended model with the low job-risk group only showed significant effects for risk aversion and negotiation experience. The goodness of fit of the extended model for the low job-risk group was worse, as indicted by a smaller adj-R^2 though.

Table 4.4 Extended model with full sample, low and high job-risk groups

Model variables	Full sample	Low job risk	High job risk
Risk informedness	-0.256*** (0.027)	-0.228 (0.045)	-0.399*** (0.042)
Risk aversion	-0.436*** (0.002)	-0.491** (0.004)	-0.581*** (0.004)
Risk informedness × Risk aversion	0.220* (0.003)	0.342 (0.006)	0.320 (0.005)
Working experience	0.064 (0.013)	0.041 (0.023)	-0.096 (0.021)
Business decision-making experience	0.018 (0.024)	0.027 (0.040)	-0.046 (-0.038)
Cloud usage experience	0.141 (0.019)	0.196 (0.032)	0.231 (0.030)
Business negotiation experience	-0.219** (0.015)	-0.306* (0.026)	-0.176 (0.024)
Business analytics experience	-0.099 (0.027)	-0.025 (0.046)	-0.170 (0.043)
Adj.-R^2	18.2%	12.2%	34%

Note: Standard deviations are reported underneath the coefficient estimates. Coefficients are standardized. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

To further examine the interaction effect between risk aversion and risk informedness, I implemented a two-sample *t*-test and a simple slope analysis. The two-sample *t*-test tests whether there were differences in the effect of risk informedness on client willingness-to-pay for subjects with low and high risk aversion. The simple slope analysis enabled me to make a visual representation of the interaction effect between risk aversion and risk informedness.

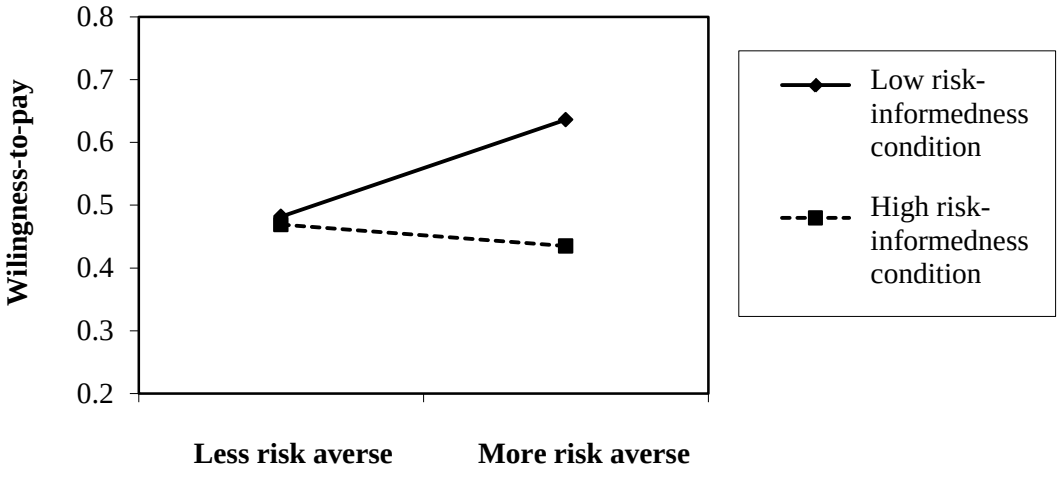
The result of the two-sample *t*-test was significant at the 0.05 level ($t = 3.50$, $p = 0.001$) for subjects with high risk aversion. In contrast, for subjects with low risk aversion (with a risk score greater than 4, and half of one standard deviation higher than the mean), the difference was not significant ($t = 0.22$, $p = 0.83$). I repeated the same test for all of the job duration conditions. (See Table 4.5.) No difference in client willingness-to-pay was found for experimental Job 1. This may have been caused by the fact that subjects were still in the process of digesting the experimental information and learning how to make an offer to the hypothetical cloud services vendor. The results for Job 2 and Job 3 were consistent.

Table 4.5 Difference in effect of risk informedness on client willingness-to-pay in the presence of different levels of job risk and client risk aversion

Subject group	Job 1 (3 hours)	Job 2 (5 hours)	Job 3 (10 hours)
Low risk aversion	0.011 (0.183)	-0.008 (0.146)	0.030 (0.135)
High risk aversion	0.132 (0.188)	0.160* (0.138)	0.230** (0.155)
Note: The difference between mean client willingness-to-pay in the baseline group and in the treatment group is reported. Standard deviations are reported in the parentheses. Signif. = *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.			

Figure 4.4 shows the results of the simple slope analysis. It presents the interaction effect between risk aversion and risk analysis support on the resulting level of client willingness-to-pay for the 5-hour jobs. A similar result is obtained for the 10-hour jobs. The results offer empirical evidence for the Client Risk Informedness Hypothesis (H3). More risk-averse clients with a lower level of risk-informedness are willing to pay more for cloud computing services with the job completion guarantee and compound pricing, based on their assessment of services interruption risk. Without a sufficient level of risk-informedness, clients will over-estimate the cost of using spot-price services, based on their downward bias for the expected value they ascribe to the outcome. Since the spot price will not drop below the base-level, this base-level price represents the certain part of the costs to the clients. Without the help of analytical tools to support rational decision-making, clients are likely to use biased heuristics and rules-of-thumb to estimate the costs associated with services interruption. As a result, some will have difficulty making confident cost estimates.

Figure 4.4 Interaction effect between risk informedness and risk aversion



I operationalized client risk-informedness based on whether subjects have access

to a risk analysis support tool. The tool helped them to achieve more accurate estimates of the likelihood of services interruption. This reduced the effects of uncertainty on their overall perceived costs. My approach was intended to improve the client's level of risk informedness.

According to the results of the post-experiment questionnaire, shown in Table 4.6, subjects from different treatment groups perceived no difference in the ease of use of the testbed, but subjects from the baseline group expressed a stronger intention to use the testbed again than subjects from the treatment group. These results indicate that although it is not difficult to use, the risk analysis support tool nevertheless introduced a considerable amount of information load on the subjects. This complicated the purchase process and lowered their intention to use the testbed again. A cloud computing services vendor or broker can develop more sophisticated support tools as value-added services to support client decision-making for using cloud computing services. It can give simple support without too much effort from clients. This also may allow the broker, as a middleman, to extract additional information rent from the client.

Table 4.6 Two-sample *t*-test results of satisfaction instruments

Instruments	Mean		Std.Dev	*p*-value
	Baseline Group	Treatment Group		
Satisfaction	4.857	4.458	1.126	0.239
Ease of use	3.810	4.125	1.462	0.473
Intention to use again	5.810	5.000	1.154	0.025**
Helpfulness of the risk analysis support tool	4.762	4.875	1.125	0.737
Risk concern vs. Profit concern	4.762	3.833	1.593	0.056*

Note: For the instrument "Risk concern vs. Profit concern", a larger number indicates that the subject was more concerned with minimizing risk than maximizing profit. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Recall that the results suggest that client willingness-to-pay for cloud computing services with guaranteed job completion under the compound pricing mechanism does not differ for job duration. It is possible that the difference in the levels of risk for spot-price services interruption for the three different duration jobs was not large enough for the subjects to have perceived meaningful differences. Alternatively, subjects may have viewed the levels of risks of these different computing jobs to be a less critical driver of their valuation of compound cloud computing services with guaranteed job completion.

4.5 Discussions and Implications

In this study, I found evidence that subjects over-estimated the risk of services interruption associated with spot-price services. As a result, they were willing to pay a higher price than actually was needed to eliminate the risk. This was more obvious when subjects were not provided with risk analysis support information. The prices offered to the hypothetical cloud services vendor were, on average, much higher than

the expected spot prices for the same computing jobs. This result confirms prior findings that consumers tend to over-estimate the probability of an undesirable outcome (Kahneman and Tversky 1979).

My results show that risk-informedness affects client willingness-to-pay for brokered cloud computing services with a job completion guarantee. This is managerially useful knowledge. My results also show that this effect, however, is moderated by the client's risk aversion. This is interesting and also is consistent with findings in previous research. In particular, Slovic (1972b) found that there was a moderating effect of subjects' intrinsic risk propensity on their biased estimation of an uncertain outcome without explicit consideration of the business problem. This study contributes empirical evidence of that finding in an IT services-related business setting.

In addition, prior research has shown that clients' aversion to risk is affected by environmental factors (Slovic 1972a). Although I controlled for environmental effects thoroughly in this experiment, it is possible and potentially interesting to extend the experiment to incorporate decision-making scenarios with different environmental settings. For example, subjects could be instructed to purchase cloud computing services for different purposes other than simulation, and under different preference settings, as well as different budget constraints.

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 (Paleologo 2004). To the spot-price services vendor, the results of this study suggest that providing additional risk analysis support will reduce users' perceived risk from services interruption and thus increase their valuation of spot-price services. This will also likely in-

crease their usage of spot-price services. Cloud services vendors therefore may want to maintain or start to offer the spot-price services, meanwhile providing risk analysis support to clients.

Cloud service brokerages are still in the early stage of development. From a cloud services broker's perspective, providing risk analysis support may not be a desirable option though, because it will lower client willingness-to-pay for the brokered cloud services. However, when risk analysis support is not available, there is an opportunity for price discrimination based on client risk aversion.

This study also provides a novel design for brokered cloud services. A services vendor can offer clients the opportunity to customize their uptime requirements, set compensation requirements, and yield the prices they are willing to pay to the vendor. The mechanism is not restricted to the customization of uptime requirements and a compensation scheme. It can also be extended to accommodate other features in the use of cloud computing services, such as service specifications, contract flexibility, and support.

There is strong market potential for cloud brokers. Without the overhead of managing and maintaining a large infrastructure and different software stacks, cloud brokers will be able to focus on innovative ways of providing services, develop a highly customized portfolios of services, and maintain their customer relationships. In addition, cloud brokers will be able to aggregate demand and supply simultaneously, which creates opportunities for better and more complex resources allocation strategies.

4.6 Conclusion

This research proposes a compound pricing mechanism for cloud computing services with guaranteed job completion. I combined an analysis of the spot prices of Amazon EC2 with an experimental study to examine the impact of key variables on client willingness-to-pay for cloud computing services sold with this new mechanism. Risk informedness, risk aversion and their interaction affect a client's willingness-to-pay. In addition, the results are consistent for jobs involving low and high completion risk.

The results suggest that increased risk informedness reduces a client's uncertainty about the impacts of services interruption. However, this informedness effect is moderated by a client's aversion to risk. The greater the client's risk aversion, the more influential will be the risk-related information. A limitation of the current experimental design is that it does not yet reveal the effects of risk-informedness in terms of whether a client is more or less prone to adopt and benefit from the compound pricing mechanism. It may be attractive for a client who is risk-averse to use guaranteed job completion services to neutralize the risk that arises around job duration. On the other hand, additional risk information may mean that a client is informed well enough to directly assess whether there is business value to be achieved from paying for protection. Currently, there are no cloud computing services vendors who offer high levels of risk analysis support so clients will be informed in the way described here. This mechanism can also be extended to accommodate other innovations in service specifications, contract flexibility and support services.

This research has limitations. Participants were required to purchase computing

resources from a hypothetical cloud vendor. I did not measure participants' actual willingness-to-pay, but instead only their hypothetical willingness-to-pay. This may cause *hypothetical bias*, which is defined in the economics literature as the bias induced by the hypothetical nature of the tasks (Harrison and Rutstrom 2008). If the testbed used in this study was deployed in a real-world business setting, incentive-aligned measurements of client willingness-to-pay will be obtained and the hypothetical bias will be overcome. Furthermore, the current version of the testbed includes only two conditions that are different in their impacts on clients' informedness about the interruption risk of the cloud services sold on this testbed. This, to a large extent, limits the usability and generality of the findings. Further development of the testbed will include the construction of a configurable module that allows modification of other desirable features related to the use of cloud services, such as whether there is protection against services interruption available, or whether a data migration tool is provided.

5 Institutional Review Board (IRB) Experience

Because the cloud computing pricing experiment conducted in Chapter 4 of this dissertation involves human subjects, prior to the conduct of this experiment, I applied for IRB approval, and gained some useful experience from the application process. In this chapter, I will provide a brief introduction to IRB, how the application proceeds, the lessons learned, and the best practices.

5.1 Introduction of IRB

The Institutional Review Board (IRB), also known as Independent Ethics Committee or Ethical Review Board, is a committee that is dedicated to the reviewing, approving, and monitoring of research projects that involve human subjects. The major responsibility of the IRB is to protect the investigators and participants. Participants may be exposed to potential risks of harm because of seemingly normal actions, and an investigator may be unaware when this may occur. Furthermore, if the information that can identify a participant is required by the investigator, the participants' privacy is at risk. Sensitive personal information also may be leaked by the investigator unintentionally. This will cause some undesirable consequences on the investigator too.

IRB committee members typically analyze the study that is being reviewed, and decide whether the study should be approved for execution. For experiments that may cause considerable risk to participants, IRB committee members will inspect its development. According to the terms and reference of SMU's IRB[2], "*The mission of the IRB is to protect the rights, privacy, and welfare of human subjects who participate in research. To accomplish this mission, the IRB will perform the primary functions of*

[2] More information about SMU IRB and IRB application can be found at http://research.smu.edu.sg/researchsmu/institutional-review-board

education, review, approval, and monitoring with regard to adherence to established criteria of ethical practices in research."

In general, IRB committee members need to make sure no sensitive personal information on the participants is collected by the investigator, unless it is absolutely necessary. In that case, they will make sure that sufficient effort is devoted, so the information is secure. IRB committee members will also examine each step of the experiment to make sure participants are not exposed to risk without sufficient justifications.

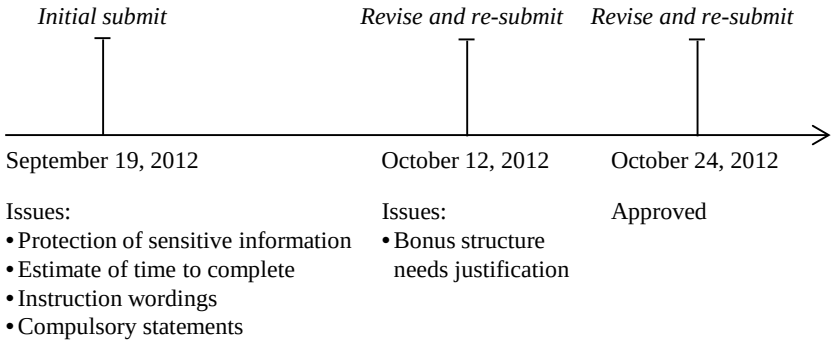
There are three categories of IRB applications: exempt from detailed review, expedited review, and full review[3]. IRB committee member will judge which category an application belongs to. Major criteria are whether participants' sensitive personal information is collected, and whether participants are exposed to risk, including deception, stress, and physical harm, etc.

5.2 Replay of IRB Permission Process

I submitted, revised, and re-submitted the IRB application for the user study "Cloud Computing Pricing Mechanisms" several times. (See Figure 5.1.) The IRB Committee members paid much attention to the details of the study and were very considerate. Detailed descriptions of the initial submission, revisions and issues raised by IRB reviewers are presented next.

[3] Detailed description of the three categories and their inclusion criteria can be found at http://intranet.smu.edu.sg/or/IRB/instructions.asp (Need to login to SMU Intranet)

Figure 5.1 Timeline for IRB application submissions and revisions



My initial submission was made on September 19, 2012. The issues were as follows:

(1) How email addresses of participants would be obtained was not specified. This needed to be stated in detail, and also who would send out email and who would be approved for accessing the emails.

(2) An estimate of the overall experiment duration, as well as separate estimates of durations of all experiment steps, should be made. In the initial submission, only an estimate of overall experiment duration was provided.

(3) Whether participants needed to provide their personal information to claim compensation was not clearly specified.

(4) One item in the application form was not completed.

(5) An improper word was used in the informed consent form ("win"), and IRB suggested a correction ("gifts of appreciation").

(6) It was not clearly specified how participants' personal information in the informed consent form would be secured.

(7) No proper statements that let participants declare they were above 18 years old were placed in the informed consent form.

(8) Contact details of the SMU IRB needed to be put in the informed consent form; they were excluded in the initial submission.

(9) A checklist for submission of IRB application needed to be submitted together with the other documents.

After receiving the comments, I carefully considered each question and revised the application forms accordingly. In this round of revision, I considered different ways of securing my subjects' sensitive information. Finally, I chose to save the sensitive data and other data separately. All the compulsory statements were available on IRB web site; some rewording is needed though.

My second submission was made on October 12, 2012. Besides the base compensation for participation, a bonus schedule was also introduced into the experiment. This is not a common component in this kind of study. I had not yet provided justification for this arrangement. The IRB required clarification on whether this bonus structure represented amounts at or close to the minimum amounts that was thought to be necessary for testing the research question. The review indicated that the total compensation given to participants was probably too much, and that it would potentially give participants an inappropriate incentive to join the experiment, and which were likely to compromise the results. So I visited the web portal of the Ministry of Manpower of Singapore to search for statistics of wages in different industries. The information that I found was used to prove that even the highest amount in the bonus structure was below the average wage level of the industry sector to which the participants belonged. This suggested that the scheme was a reasonable arrangement.

The third submission was on October 24, 2012, and was approved on October 29,

2012. All documents had to be submitted in hard copy and needed to be signed by both the primary investigator and the applicant.

5.3 Lessons Learned from the IRB Application Procedure

During the IRB application, I went through the whole experimental process more times than I expected. It encouraged me to keep asking questions about each step of the experiment: (1) is it necessary to answer the research question? (2) Will it cause any risk to the subjects? (3) How long will the experiment take, and will it be too time consuming? Through this process, I found that several steps actually could be dropped and some could be combined, thus simplifying the experimental procedure.

In addition, the IRB application increased my awareness of the necessity of securing the subjects' personal information. I also learned the importance of establishing a proper compensation scheme for participations. This led me to refine the incentive design.

5.4 Best Practices

Despite the kind help from the IRB staff and detailed instructions on how to develop an application, there are still some issues to which an applicant should pay special attention in order to improve the quality of the research and reduce the rounds of revisions.

- *Pay extra attention to the details*. It's always a good practice to prepare a checklist at the beginning of the application process, when you start to prepare the related documents. Besides the official checklist that identifies all of the documents that are to be submitted, it's helpful to list the key components in the experiment design too. That will include: (1) a clearly written statement on

the motivation and contribution of the research; (2) a deception of reasonable compensation scheme; (3) a description of the steps that involve collection of subjects' personal information and how the information will be secured; (4) estimation of the time a subject will spend in each step of the experiment, and an estimate of the time to be spent for the whole experiment; (5) IRB contact details in the informed consent form; (6) a proper statement in the informed consent form that requires participants to declare they are above 18 years old; and (7) a statement that helps participants to be aware of the fact that they are participating the experiment voluntarily and that they are free to discontinue participation anytime without penalty.

- *Design an effective incentive scheme*. It's very important to properly motivate subjects to participate. The background and research questions of the study should be conveyed to potential participants in a clear way. Highlighting their contributions in plain English will be helpful in attracting potential participants who will be interested in the research. Compensation should be set up properly too. It should be no higher than the average hourly wage level of the targeted subject group times the estimated duration of the whole experiment in hours. It is a token to show the experimenter's appreciation to the participants for spending their time and energy contributing to the study. But it's not meant to be a way to incentivize subjects to participate in the experiment.
- *Try not to store sensitive personal information*. I am referring to sensitive personal information addresses that can potentially identify a specific person, such as an email address with real names, or a passport number. Think twice

while designing the experimental study about whether it's a necessity to collect such information. If the information is used just to identify subjects in order to give them compensation, then masking identifiers, such as a sequence of randomly generated integers, can be used instead. If it's necessary to collect sensitive personal information, store this information and other experimental data separately and with security treatments such as data encryption.

- *Give simple, clear, and correct instructions*. Subjects commonly are required to read and follow instructions to complete an experiment. The instructions should be correct and clear. Using accurate descriptions and eliminating ambiguous ones are important.
- *Double check the experimental procedures for potential deception and risk.* Subjects commonly are required to perform some tasks during the experimental procedures. The experimental tasks, procedures, as well as instruction, should not introduce any deception to subjects. All steps should be carefully examined to ensure that they are necessary to seeking answers to the research questions, and won't cause potential physical or mental damage to subjects.

6 Conclusion, Limitations and Future Research

Chapter 2 provided an overview of the macro-structure of pricing practices in the cloud computing services market. I conducted a market survey of pricing mechanisms implemented by representative cloud computing services vendors. Based on the information I extracted the structure of price-related attributes, and created a framework of pricing models for cloud computing services. The following two chapters adopted micro-level perspectives to study pricing in the cloud computing services market. One takes the vendor's point of view and the other takes the client's perspective.

Chapter 2 explored in detail the pricing schema of 27 cloud computing services from 19 representative cloud services vendors. I found the commonly adopted usage based pricing is decided not only by the metered usage, but also by whether a client needs to be assured of access to the services, how much the client would expect if the services were down for a certain period of time, and how much support the client needed from the cloud services vendor. I also found the cost of using cloud computing services decreased along time, with frequent price reductions announced by cloud services vendors. In addition, clients have more flexibility in choosing their ideal combination of cloud resources to come up with a configuration that serve their needs best. This is enabled by cloud services vendors' effort to make their services offerings configurable, and prices for customized configurations instantly shown.

Chapter 3 investigated a cloud vendors' pricing strategy in the presence of fixed-price reserved services and spot-price on-demand services. Hybrid pricing strategy, involving both fixed-prices and spot-prices, is preferable for a cloud services vendor with market power, when clients are sensitive to services interruption or a lower qual-

ity version of the cloud services yields reasonably high profit for the vendor. The interplay between services offered with the two pricing methods and their impacts on vendor profit and client welfare is also discussed. By introducing spot-price interruptible services of lower quality, the vendor gains a higher profit, while the clients' welfare decreases.

Chapter 4 takes the client perspective to investigate factors affecting their willingness-to-pay for brokered cloud computing services that permit some level of customization. Specifically, the customization of cloud computing services that I consider is related to the level of risk of services interruption. I did experimental work involving hypothetical cloud computing services clients, with two design variables: risk-informedness and task duration. *Task duration* is a proxy for the level of risk for a task to be executed using spot-price cloud services. The results indicated a significant effect of *risk infromedness* on client willingness-to-pay. They also suggested an interaction effect between risk informedness and client risk aversion. So it is important for vendors to gain knowledge about their clients' risk aversion to achieve higher profit. Without such knowledge, vendors will not be able to implement pricing plans that allow clients to self-select according to their risk aversion.

The studies in this thesis research have limitations. For the analytic model, spot prices are assumed to be exogenous. A more general setting should consider a vendor's decision on spot prices together with its decision on the fixed price of reserved services. In addition, I assumed that clients' demands are invariant across all periods, which is not realistic. Furthermore, the pricing experiment did not involve participants that were participating in the real-world cloud computing business. The experi-

mental study can be well extended to include subjects who are in real cloud businesses.

My future research will include a relaxation of the i.i.d. assumption in the analytic model that addresses cloud computing pricing strategy. I also plan to relax the exogenous spot price level assumption, and mimic the real Amazon EC2 market with a simulation approach. I have refinement of the testbed for more configurable cloud pricing experiment settings in progress. The risk analysis approach also needs improvement, in order to permit predictive analysis support that is a desirable feature of clients. The pricing model will also support the comparison of services that are offered by direct competitors. I will also extend the pricing model to include the consideration on how price signals quality and resource capacity of cloud computing services, and provide predictions related to future shifts in services specifications and pricing approaches.

Bibliography

Aedagna, D., Panicucci, B., and Passacantando, M. (2013). Generalized Nash equilibria for the service provisioning problem in cloud systems. Forthcoming in *IEEE Transactions on Services Computing*.

Amazon (2009). Announcing Amazon EC2 Spot Instances. Accessed on October 9, 2013. Available at http://aws.amazon.com/about-aws/whats-new/2009/12/14/announcing-amazon-ec2-spot-instances/

Amazon EC2 Instances. Accessed on October 5th, 2013. Available at: http://aws.amazon.com/ec2/instance-types/

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50-58.

Bapna, R., Dellarocas, C., and Rice. S. (2010). Vertically differentiated simultaneous Vickrey auctions: theory and experimental evidence. *Management Science*, 56 (7), 1074-1092.

Bardhan, I. R., Demirkan, H., Kannan, P. K., Kauffman, R. J., and Sougstad, R. (2010). An interdisciplinary perspective on IT services management and service science. *Journal of Management Information Systems*, 26(4), 13-64.

Benaroch, M., Dai, Q., and Kauffman, R. J. (2010). Should we go our own way? Backsourcing flexibility in IT services contracts. *Journal of Management Information Systems*, 26(4), 317-358.

Bhargava, H. K. and Choudhary, V. (2008). Research note, When is versioning optimal for information goods? *Management Science*, 54(5), 1029-1035.

Boiteux, M. (1960). Peak-load pricing. *The Journal of Business*, 33(2), 157-179.

Boyd, E. A., and Bilegan, I. C. (1999). Revenue management and e-commerce. *Management Science*, 49(10), 1363-1386.

Brandon. (2011) Amazon EC2 spot request volatility hits $1,000/hour. MozDev, September 27.

Broberg, J., Venugopal, S., and Buyya, R. (2008). Market-oriented grids and utility computing: The state-of-the-art and future directions. *Journal of Grid Computing*, 6(3), 255-276.

Buyya, R., Abramson, D., and Venugopal, S. (2005). The grid economy. *Proceedings of the IEEE*, 93(3), 698-714.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., and Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6), 599-616.

Chaisiri, S., Lee, B., and Niyato, D. (2012). Optimization of resource provisioning cost in cloud computing. *IEEE Transactions on Services Computing*, 5(2), 164-177.

Chen, P. Y., and Wu, S. Y. (2012). The impact and implications of on-demand services on market structure. *Information Systems Research*, 24(3), 750-767.

Choudhary, V. (2010). Use of pricing schemes for differentiating information goods. *Management Science*, 21(1), 78-92.

Choudhary, V. (2010). Use of pricing schemes for differentiating information goods. *Information Systems Research*, 21(1), 78-92.

Clemons, E.K. and Gao, G. (2008). Consumer informedness and diverse consumer purchasing behaviors: traditional mass-market, trading down, and trading out into the long tail. *Electronic Commerce Research and Applications*, 7 (1), 3-17.

Dana, J.D., Jr. (1999). Using yield management to shift demand when the peak time is unknown. *RAND Journal of Economics*, 30(3), 456-474.

Demirkan, H., Kauffman, R. J., Vayghan, J. A., Fill, H. G., Karagiannis, D., and Maglio, P. P. (2009). Service-oriented technology and management: Perspectives on research and practice for the coming decade. *Electronic Commerce Research and Applications*, 7(4), 356-376.

Deneckere, R. J. and McAfee, R. P. (1996). Damaged goods. *Journal of Economics and Management Strategy*, 5(2), 149-174.

Etro, F. (2009). The economic impact of cloud computing on business creation, employment and output in Europe. *Review of Business and Economics*, 54(2), 179-208.

Fishburn, P. C. (1988). Expected utility: an anniversary and a new era. *Journal of Risk and Uncertainty*, 1(3), 267-283.

Fishburn, P. C., and Odlyzko, A. M. (1999). Competitive pricing of information goods: Subscription pricing versus pay-per-use. *Economic Theory*, 13(2), 447-470.

Fishburn, P. C., Andrew, M. O., and Ryan, C. S. (2000). Fixed fee versus unit pricing for information goods: competition, equilibria, and price wars. In D. Hurley, B. Kahin, and H. Varian (eds.), *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property*, 167-189, MIT Press, Boston, MA.

Foster, I., Zhao, Y., Raicu, I., and Lu, S. (2008). Cloud computing and grid computing 360-degree compared. In *Grid Computing Environments Workshop 2008, GCE '08*, November 12-16, 2008, Austin, TX.

Frost, V. S., and Melamed, B. (1994). Traffic modeling for telecommunications networks. *Communications Magazine, IEEE*, 32(3), 70-81.

Gallego, G., and Van Ryzin, G. (1997). A multiproduct dynamic pricing problem and its applications to network yield management. *Operations Research*, 45(1), 24-41.

Gartner. (2011). Cloud computing: cloud services brokerage. Report, Stamford, CT.

Gartner (2012). Gartner says worldwide cloud services market to surpass $109 billion in 2012. Stamford, CT.

Gartner (2013). Gartner says worldwide public cloud services market to total $131 billion. Stamford, CT.

Ghosh, R., Longo, F., Xia, R., Naik, V. K., and Trivedi, K. S. (2013). Stochastic model driven capacity planning for an infrastructure-as-a-service cloud. Forthcoming in *IEEE Transactions on Services Computing*.

Gregory, R., Lichtenstein, S., and Slovic, P. (1993). Valuing environmental resources: A constructive approach. *Journal of Risk Uncertainty*, 7(2), 177–97.

Grindle, M., Kavathekar, J., and Wan, D. (2013). A new era for the healthcare industry. Accenture report, Yew York, NY.

Harrison, G. W. and Rutström, E. E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods, in *Handbook of Experimental Economics Results*, Charles R. Plott and Vernon L. Smith, eds., Amsterdam, North-Holland, 727–67.

Heath, N. 2013. AWS lowers prices for the 30th time, but do customers really care anymore? Accessed on October 09, 2013. Available from: http://www.zdnet.com/aws-lowers-prices-for-the-30th-time-but-do-customers-really-care-anymore-7000013530/

Henzinger, T. A., Singh, A. V., Singh, V., Wies, T., and Zufferey, D. (2010, October). A marketplace for cloud resources. In *Proceedings of the 10th ACM International Conference on Embedded Software, EMSOFT 2010,* Scottsdale, AZ.

Higginbotham, S. (2009) Dynamic pricing comes to Amazon's cloud. GigaOM, San Francisco, CA, December 14.

Howard, B. (2011). Crawl outage – An update and what we're doing. The Moz Blog. Accessed on October 9, 2013. Available from: www.seomoz.org/blog/crawl-outage.

Hwang, R., Lee, C., Chen, Y., and Zhang-Jian D. (2013). Cost optimization of elasticity cloud resource subscription policy. Forthcoming in *IEEE Transactions on Services Computing*.

Iosup, A., Li, H., Jan, M., Anoep, S., Dumitrescu, C., Wolters, L., and Epema, D.H.J. (2008). The Grid Workloads Archive. *Future Generation Computer Systems*, 24 (7), 672-686.

Jackson, K. (2012). Cloud management broker: The next wave in cloud computing. *Forbes*, October 12, 2012.

Javadi, B., Thulasiramy, R. K., and Buyya, R. (2011). Statistical modeling of spot instance prices in public cloud environments. In *4th IEEE International Conference on Utility and Cloud Computing* (UCC 2011), December 5-7, 2011, Mel-

bourne, Australia.

Jayasinghe, D., Malkowski, S., Li, J., Wang, Q., Wang, Z., and Pu, C. (2013). Variations in performance and scalability: An experimental study in IaaS clouds using multi-tier workloads. Forthcoming in *IEEE Transactions on Services Computing*.

Kesavan, M., Ahmad, I., Krieger, O., Soundararajan, R., Gavrilovska, A., and Schwan, K. (2013). Practical compute capacity management for virtualized datacenters. Forthcoming in *IEEE Transactions on Cloud Computing*.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47 (2), 263-291.

Kondo, D., Javadi, B., Malecot, P., Cappello, F., and Anderson, D. P. (2009). Cost-benefit analysis of cloud computing versus desktop grids. In *Proceedings of IEEE International Symposium on Parallel & Distributed Processing, IPDPS 2009*. IEEE, Rome, Italy, pp.1-12.

Krupnick, A., Alberini, A., Cropper, M., Simon, N., O'Brien, B., Goeree, R., and Heintzelman, M. (2002). Age, health and the willingness to pay for mortality risk reductions: A contingent valuation survey of Ontario residents. *Journal of Risk and Uncertainty*, 24(2), 161-186.

Leong, L. and Chamberlin, T. (2010). Magic quadrant for cloud infrastructure as a service and web hosting. Gartner Research Report. Stamford , CT.

Leong, L. and Chamberlin, T. (2011). Magic quadrant for public cloud infrastructure as a service. Gartner Research Report. Stamford , CT.

Leong, L., Toombs, D., Gill, B., Petri, G., and Haynes, T. (2012). Magic quadrant for cloud infrastructure as a service.Gartner Research Report. Stamford , CT.

Leong, L., Toombs, D., Gill, B., Petri, G., and Haynes, T. (2013). Magic quadrant for cloud infrastructure as a service. Gartner Research Report. Stamford , CT.

Lheureux, B.J. and Plummer, D.C. (2012). Cloud services brokerages: A high-impact opportunity for IT. Gartner Research Report.

Li, H., Muskulus, M., and Wolters, L. (2007). Modeling job arrivals in a data-intensive grid. In *Job Scheduling Strategies for Parallel Processing*. Springer Berlin Heidelberg, pp.210-231.

Li, M. Z. F. (1994). Pricing perishable inventories by using marketing restrictions with applications to airlines. Doctoral dissertation, University of British Columbia, Vancouver, Canada.

Li, X., Guang, T. H. G., and Veeravalli, B. (2006). Design and implementation of a multimedia personalized service over large scale networks. In *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo* (*ICME 2006*), July 9-12, 2006, Toronto, Ontario, Canada. IEEE, pp.77-80.

Li, X., Hsieh, J. J. P., and Rai, A. (2013). Motivational differences across post-acceptance information system usage behaviors: An investigation in the business intelligence systems context. *Information Systems Research*, 24(3), 659-682, 879-881.

Ma, D. and Seidmann, A. (2008). The pricing strategy analysis for the "software-as-a-service" business model. In *Grid Economics and Business Models*. Springer, Berlin and Heidelberg, pp.103-112.

Madhavaiah, C., Bashir, I., and Shafi, S. I. (2012). Defining cloud computing in business perspective: A review of research. *Vision: The Journal of Business Perspec-*

tive, 16(3), 163-173.

Malone, T., Yates, J., and Benjamin, R. (1987). Electronic markets and electronic hierarchies. *Communications of the ACM*, 30 (6), 484-497.

Markopoulos, P.M. and Clemons, E. (2013). Reducing Buyers' Uncertainty about Taste-Related Product Attributes. Forthcoming in *Journal of Management Information Systems*.

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., and Ghalsasi, A. (2011). Cloud computing—The business perspective. *Decision Support Systems*, 51(1), 176-189.

Maskin, E., and Riley, J. (1984). Monopoly with incomplete information. *RAND Journal of Economics*, 15(2), 171-196.

Masuda, Y. and Whang, S. (2006). On the optimality of fixed-up-to tariff for telecommunications service. *Information Systems Research*, 17(3), 247-253.

Mei, Y., Liu, L., Pu, X., Sivathanu, S., and Dong, X. (2011). Performance analysis of network I/O workloads in virtualized data centers. *IEEE Transactions on Services Computing*, 6(1), 48-63.

McAfee, R. P. (2007). Pricing damaged goods. *Economics: The Open-Access, Open-Assessment E-Journal*, 1(1), 1-19.

McGill, J. I. and Van Ryzin, G. J. (1999). Revenue management: Research overview and prospects. *Transportation Science*, 33(2), 233-256.

Miller, K. M., Hofstetter, R., Krohmer, H., and Zhang, Z. J. (2011). How should consumers' willingness to pay be measured? An empirical comparison of state-of-the-art approaches. *Journal of Marketing Research*, 48(1), 172-184.

Moorthy, K. S. and Png, I. P. (1992). Market segmentation, cannibalization, and the

timing of product introductions. *Management Science*, 38(3), 345-359.

Ou, Z., Zhuang, H., Lukyanenko, A., Nurminen, J. K., Hui, P., Mazalov, V., and Yia-Jaaski, A. (2013). Is the same instance type created equal? Exploiting heterogeneity of public clouds. Forthcoming in *IEEE Transactions on Cloud Computing*.

Paleologo, G. A. (2004). Price-at-risk: A methodology for pricing utility computing services. *IBM Systems Journal*, 43(1), 20-31.

Parolini, L., Tolia, N., Sinopoli, B., and Krogh, B. H. (2010). A cyber-physical systems approach to energy management in data centers. In *Proceedings of the 1st ACM/IEEE International Conference on Cyber-Physical Systems* (*ICCPS*), April 13-14, 2010, Stockholm, Sweden. ACM, pp.168-177.

Png, I. P. L., and Wang, H. (2010). Buyer uncertainty and two-part pricing: Theory and applications. *Management Science*, 56(2), 334-342.

Pu, X., Liu, L., Mei, Y., Sivathanu, S., Koh, Y., Pu, C., and Cao, Y. (2013). Who is your neighbor: Net I/O performance interference in virtualized clouds. *IEEE Transactions on Services Computing*, 6(3), 314-329.

Relihan III, W. J. (1989). The yield-management approach to hotel-room pricing. *Cornell Hotel and Restaurant Administration Quarterly*, 30(1), 40-45.

Rimal, B. P., Choi, E., and Lumb, I. (2009). A taxonomy and survey of cloud computing systems. In *5th International Joint Conference on INC, IMS and IDC* (*NCM '09*). IEEE, Seoul, Korea, pp.44-51.

Schmidt, U., Starmer, C., and Sugden, R. (2008). Third-generation prospect theory. *Journal of Risk and Uncertainty*, 36 (3), 203-223.

Schoemaker, P. J. H. (1993). Determinants of risk-taking: Behavioral and economic

view. *Journal of Risk and Uncertainty*, 6 (1), 49-73.

Sim, K. M. (2010). Towards complex negotiation for cloud economy. In *Advances in Grid and Pervasive Computing*. Springer, Berlin and Heidelberg, pp.395-406.

Slovic, P. (1972a). Information processing, situation specificity, and the generality of risk-taking behavior. *Journal of Personality and Social Psychology*, 22 (1), 128-134.

Slovic, P. (1972b). Psychological study of human judgmental implications for investment decision making. *Journal of Finance*, 27(4), 779-799.

Smith, G. E. and Nagle, T. T. (1995). Frames of reference and buyers' perception of price and value. *California Management Review*, 38(1), 98-116.

Smith, V.L. (1976). Experimental economics: Induced value theory. *American Economic Review*, 66(2), 274–279.

Speitkamp, B. and Bichler, M. (2010). A mathematical programming approach for server consolidation problems in virtualized data centers. *IEEE Transactions on Services Computing*, 3(4), 266-278.

Sridhar, B., Bhattacharya, S., and Krishnan, V. (2011). Pricing information goods: A strategic analysis of the selling and on-demand pricing mechanisms. Working paper, Smeal College of Business, Pennsylvania State University, College Park, PA.

Steiner, P. O. (1957). Peak loads and efficient pricing. *Quarterly Journal of Economics*, 71(4), 585-610.

Stößer, J., Neumann, D., and Weinhardt, C. (2010). Market-based pricing in grids: On strategic manipulation and computational cost. *European Journal of Operational Research*, 203(2), 464-475.

Sundararajan, A. (2004). Nonlinear pricing of information goods. *Management Science*, 50(12), 1660-1673.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.

Urbany, J. E., Dickson, P. R., and Wilkie, W. L. (1989). Buyer uncertainty and information search. *Journal of Consumer Research*, 16(2), 208-215.

Vaquero, L. M., Rodero-Merino, L., Caceres, J., and Lindner, M. (2008). A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review*, 39(1), 50-55.

Varian, H. R. (1995). Pricing information goods. In *Proceedings of Scholarship in the New Information Environment Symposium.* Harvard Law School, Boston, MA.

Varian, H.R. (1997). Versioning information goods. Working paper, University of California, Berkeley, CA.

Wang, T., Venkatesh, R., and Chatterjee, R. (2007). Reservation price as a range: An incentive-compatible measurement approach. *Journal of Marketing Research*, 44(2), 200-213.

Weber, E.U., Blais, A.R., and Betz, N.E. (2002). A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15 (4), 263-290.

Williamson, E. L. (1992). Airline network seat inventory control: Methodologies and revenue impacts. Doctoral dissertation, Flight Transportation Laboratory, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA.

Wilson, G.W. (1972). The theory of peak-load pricing: a final note. *The Bell Journal of Economics and Management Science*, 3(1), 307-310.

Wu, L., Garg, K. S., Versteeg, S., and Buyya, R. (2013). SLA-based resource provisioning for hosted software as a service applications in cloud computing environments. Forthcoming in *IEEE Transactions on Services Computing*.

Yi, S., Andrzejak, A., and Kondo, D. (2012). Monetary cost-aware checkpointing and migration on Amazon cloud spot instances. *IEEE Transactions on Services Computing*, 5(4), 512-524.

Yi, S., Kondo, D., and Andrzejak, A. (2010). Reducing costs of spot instances via checkpointing in the Amazon Elastic Compute Cloud. In *3rd IEEE International Conference on Cloud Computing* (CLOUD 2010), July 5-10, Miami, Florida. IEEE, pp.236-243.

Zhao, L., Sakr, S., and Liu, A. (2013). A framework for consumer-centric SLA management of cloud-hosted databases. Forthcoming in *IEEE Transactions on Services Computing*.

Appendices

- **Appendix A. Pricing Detail of Cloud Computing Services Vendors**
- **Appendix B. Proofs for Lemmas and Propositions**
- **Appendix C. Companions of the Experimental Study**

Appendix A1. Pricing Detail of IaaS Cloud Services

	Factor Group I: Usage-based Pricing			Factor Group II: Reservation-based Pricing		Factor Group III: Support-related Pricing		Factor Group IV: Penalties	
	Services Specifications	Unit Price	Total Usage	Reservation Period	On-time Payment	Support Services Specification	Unit Price	Services Down Time	Penalty Rate
Amazon EC2 a. On-Demand Instance	- OS (Linux or Windows), - size of instances (small, large, extra large), -location of server (US, Europe, Asia)	Varies with services specifications, for example, $0.085 /hour for Linux, Small, US-N.V.	[0, Upper Bound*]	NA	NA	Bronze/Silver/ Gold/Platinum Response within 15 minutes to12 hours	Varies with services specifications and usage charge, for example, for Platinum Support, Cost will be the greater of $15K or 10% of monthly usage charge	Annual uptime for the services is fixed at 99.95% Response time guarantee for support services	A 10% service credit if the services uptime not met
Amazon EC2 b. Reserved Instances		Varies with services specifications, for example, $0.03 /hour for an instance (Linux, Small, US-N.V).		1 year/3 years	Varies with services specification and reservation period, for example, $227.5 for 1 year reservation of a Linux, Small, US-N.V. instance.				
Amazon EC2 c. Spot Instances		Varies with services specifications, for example, $0.031 /hour for Linux, Small, US-N.V.	[0, Available Resource*]	NA	NA				
Amazon S3	99.999999999% Durability / 99.99% Durability	Varies with services specifications, for example, $0.14/ GB, $0.01 per 1000 requests	[0, Upper Bound*]	NA	NA			Monthly uptime of 99.9% for the services Response time guarantee for supported services	A 10%-25% service credit if services uptime not met

Appendix A1. Pricing Detail of IaaS Cloud Services (Cont.d)

	Factor Group I: Usage-based Pricing			Factor Group II: Reservation-based Pricing		Factor Group III: Support-related Pricing		Factor Group IV: Penalties	
	Services Specifications	Unit Price	Total Usage	Reservation Period	On-time Payment	Support Services Specification	Unit Price	Services Down Time	Penalty Rate
Alatum	-Compute -ValuePlus -Power -PowerPlus -Customized	Varies with services specifications, for example, S$479.70/month for Value Compute	[0, Upper Bound*]	1 year	Info not available	Info not available	Info not available	Info not available	Info not available
nGrid	Compute	S$1.00/hour	[0, Upper Bound*]	1 month (package)	S$850	Info not available	Info not available	Info not available	Info not available
CloudSigma	-RAM -Cores -Storage	Varies with the services specifications, for example, S$72.37/month for an instance with 1 Core, 2GB memory, and 80GB SSD storage	[0, Upper Bound*]	1 month up to 3 years	Based on services specifications.	Info not available	Info not available	A 100% uptime guarantee for the services	Credit of 50 times the fees for any period: (1) virtual server unavailable; or (2) network unavailable for more than 15 minutes
GoGrid Cloud Servers	-RAM -Cores -Storage	Varies with services specifications, for example, $0.04/hour for an X-Small instance	[0, Upper Bound*]	-Monthly -Semiannually -Annually	Varies with services specifications, for example, $18.13/month for an X-Small instance	Free 24/7 tech support	Free	A 100% uptime guarantee for the services A 30 mins emergency response time for support services	10000% Service Credit if uptime not met

Appendix A1. Pricing Detail of IaaS Cloud Services (Cont.d)

	Factor Group I: Usage-based Pricing			Factor Group II: Reservation-based Pricing		Factor Group III: Support-related Pricing		Factor Group IV: Penalties	
	Services Specifications	Unit Price	Total Usage	Reservation Period	On-time Payment	Support Services Specification	Unit Price	Services Down Time	Penalty Rate
Joyent Cloud	-RAM -Cores -Storage	Varies with services specifications, for example, S$0.04/hour for an Extra Small instance	[0, Upper Bound*]	NA	NA	-Bronze -Gold -Platinum	Free, but depends on the consumption levels.	A 100% availability for the services	5% of the monthly fee for each 30 minutes of downtime (up to 100% of customer's monthly fee for the affected server)
RackSpace Cloud Server	-RAM -vCPU -Storage -Network -OS	Varies with services specifications, for example, $0.08/hour for an instance with 1 vCPU, 1GB RAM, 40GB disk space, 30Mbps public network, 60Mbps internal network	[0, Upper Bound*]	NA	NA	-Managed Cloud -Cloud Account	Different support components based on clients' account types	A 100% uptime guarantee for the services	5% of the monthly fee for each 30 minutes of downtime (up to 100% of the cloud server fees)
FlexiScale	-RAM -vCPU -Storage -Network -OS	Varies with services specifications, charged with virtual credits (purchased with cash)	[0, Upper Bound*]	NA	NA	-Phone support -Online Customer Support Ticket or email -24x7 Emergency helpline	Free	A 100% monthly uptime guarantee for the services	Credit 5% of the monthly services units for each additional 30 minutes of downtime that occur, up to 100% of the monthly services units consumed.
Profit Bricks IaaS	-RAM -Core -Storage -Traffic	Varies with services specifications, for example, $6¢/hour for 1 Core (=4 CPUs)	[0, Upper Bound*]	NA	NA	24/7 phone support	Free	A 99.95% annual uptime guarantee for the services	NA

Appendix A1. Pricing Detail of IaaS Cloud Services (Cont.d)

	Factor Group I: Usage-based Pricing			Factor Group II: Reservation-based Pricing		Factor Group III: Support-related Pricing		Factor Group IV: Penalties	
	Services Specifications	Unit Price	Total Usage	Reservation Period	On-time Payment	Support Services Specification	Unit Price	Services Down Time	Penalty Rate
Google Compute Engine	-Core -Memory -Storage	Varies with services specifications, for example, S$0.17/hour for a n1-standard-1d instance	[0, Upper Bound*]	NA	NA	-Community support -Bronze -Silver -Gold -Platinum Support	Varies with support services specifications, for example, Bronze is free, Silver costs $150 per month	A 99.95% monthly uptime guarantee for the services	-10% of monthly bill as service credit for 99.00%-99.95% -25% credit for 95.00% - < 99.00% -50% credit for below 95.00%
HP Public Cloud (Cloud Compute)	-OS (Linux, Windows) -size of compute instances: extra small, small, medium, large, extra large, double extra large	Varies with services specifications, for example, S$0.04/hour for an extra small instance with 1 HP Cloud Compute Unit (1 virtual core w/1 HP Cloud Compute Unit), 1GB RAM, 30GB storage	[0, Upper Bound*]	Month	Varies with services specifications, and number of instances	-Community forums -Customer knowledge base -Support case management page -Live support-chat, email, and phone -Status page	Free	A 99.95% monthly availability guarantee for the services	-5% of monthly bill as service credit for 99.9% to 99.95% -10% credit for 99.5% to 99.9% -20% credit for 99% to 99.5% for -30% credit for below 99.0%
CloudLayer Computing	Core, Ram, Storage, Outbound traffic (Inbound traffic free)	Varies with services specifications, for example, S$0.10/hour for an instance with 1 Core, 1GB RAM, 25GB storage	[0, Upper Bound*]	Month	Varies with services specifications, and number of instances	-Web -Phone support -Ticket system support	Free	100% network uptime guarantee, 20 minutes response on all tickets	service credit of 5% of the fees for the relevant service for every 30 continuous minutes outage period

Appendix A2. Pricing Detail of PaaS Cloud Services

	Factor Group I: Usage-based Pricing			Factor Group II: Reservation-based Pricing		Factor Group III: Support-related Pricing		Factor Group IV: Penalties	
	Services Specifications	Unit Price	Total Usage	Reservation Period	On-time Payment	Support Services Specification	Unit Price	Services Down Time	Penalty Rate
Microsoft a. Windows Azure	-OS -Size of instances (small, large, extra large) -Location (US, Europe, Asia)	Varies with services specifications, for example, S$0.06/hour for an extra small instance	[0, Upper Bound*]	1 month	Varies with services specifications, for example, S$90.35/month	-Foundation -Standard -Plus -Ultimate	Varies with support services specifications	A 99.95% monthly uptime guarantee for services A 1 hour response time guarantee for support services	A 10%-25% service credit if uptime not met
Microsoft b. SQL Azure	-Web edition -Business edition	Varies with services specifications, for example, S$12.54/month for one Web Edition database up to 1GB		1 month	Varies with services specifications, for example, S$100.40/month for a base Business Edition unit			A 99.9% monthly uptime guarantee for the services A 1 hour response time guarantee for support services	
Force.com	-One App -Enterprise -Unlimited	Info not available	[0, Upper Bound*]	1 month	Varies with service specifications, for example, S$18.83/month for One App edition of one user	-Standard -Premier -Premier+	Price depends on subscription type and monthly fee.	A 2 hours or 2 business days response time guarantee for the support services	NA

Appendix A2. Pricing Detail of PaaS Cloud Services (Cont.d)

	Factor Group I: Usage-based Pricing			Factor Group II: Reservation-based Pricing		Factor Group III: Support-related Pricing		Factor Group IV: Penalties	
	Services Specifications	Unit Price	Total Usage	Reservation Period	On-time Payment	Support Services Specification	Unit Price	Services Down Time	Penalty Rate
Google App Engine	Standard	S$11.30/app S$627.55/account	[0, Upper Bound*]			-Community Support -Operational Support -Premium Developer Support		A 99.95% monthly uptime guarantee for the services A 1 to 8 business hours response time guarantee for the support services	A 10%-50% monthly service credit if uptime not met
Amazon Beanstalk	No additional charge besides charges for EC2 instances	NA	[0, Upper Bound*]	NA	NA	NA	NA	NA	NA
CloudFare	-Free -Pro -Business -Enterprise	NA	[0, Upper Bound*]	1 month	Varies with services specifications, for example, S$251.02/month for each site under Business plan	-Email -Phone support -Dedicated account manager	Free, option availability depends on the plan	A 100% uptime guarantee for Business and Enterprise plans	Service Credit = (Outage Period minutes * Affected Customer Ratio) ÷ Scheduled Availability minutes

Appendix A3. Pricing Detail of SaaS Cloud Services

	Factor Group I: Usage-based Pricing			Factor Group II: Reservation-based Pricing		Factor Group III: Support-related Pricing		Factor Group IV: Penalties	
	Services Specifications	Unit Price	Total Usage	Reservation Period	On-time Payment	Support Services Specification	Unit Price	Services Down Time	Penalty Rate
SalesForce a. Service Cloud	-Professional -Enterprise -Unlimited	Information not available	[0, Upper Bound*]	1 month	Varies with services specifications, for example, S$81.58/user/month for Professional Edition	-Standard -Premier -Premier+	Standard Support is included in all service packages. Premier: 15% of license price for Professional and Enterprise Editions. Premier+: Included with Unlimited Edition. Or 25% of license price for Professional and Enterprise Editions.	A 2 hours or 2 business days response time guarantee for the support services	Information not available
SalesForce b. Sales Cloud	-Contact Manager -Group -Professional -Enterprise -Unlimited				Varies with services specifications, for example, S$18.83/user/month for Group Edition				
SalesForce c. Chatter	-Chatter(free) -Chatter Plus				Varies with services specifications, for example, S$18.83/user/month for Chatter Plus				
SalesForce c. Jigsaw	-Jigsaw Clean -Jigsaw -Jigsaw Lists				Varies with services specifications, for example, S$23.85 /user/month for Jigsaw Clean				

Appendix A3. Pricing Detail of SaaS Cloud Services (Cont.d)

	Factor Group I: Usage-based Pricing			Factor Group II: Reservation-based Pricing		Factor Group III: Support-related Pricing		Factor Group IV: Penalties	
	Services Specifications	Unit Price	Total Usage	Reservation Period	On-time Payment	Support Services Specification	Unit Price	Services Down Time	Penalty Rate
Google App for Business	-Flexible plan -Annual plan	NA	[0, Upper Bound*]	month / year	S$6.28/user account/month S$62.76/user account/year	-Standard -Premium	Information not available	A 99.9% monthly uptime guarantee for the services A 1 hour to 1 business day response time guarantee for the support services	3-15 days free use if services uptime not met
NetSuite Cloud ERP Software Suite	Various packages and plans for different software applications	NA	[0, Upper Bound*]	month	Varies with services specifications, and number of users, for example, S$123.9/month for NetSuite Small Business	-NetCARE Silver -Gold -Gold 24/7 -Platinum	-Silver: 22.5% of net license amount -Gold: 27.5% of net license amount -Gold 24/7: 32.5% of net license amount -Platinum: 37.5% of net license amount	A 99.5% uptime guarantee for the services	Information not available
Microsoft Office 365	-For Home -For Small Business -For Midsize Business	NA	[0, Upper Bound*]	month / year	Varies with services specifications, number of concurrent users, and number of by-service components	Varies with services specifications (blog, wiki, QA, phone, direct technical staff phone contact, etc.)	Free	Varies with services specifications	NA

Appendix A4. Pricing Detail of Cloud Brokerage Services

	Factor Group I: Usage-based Pricing			Factor Group II: Reservation-based Pricing		Factor Group III: Support-related Pricing		Factor Group IV: Penalties	
	Services Specifications	Unit Price	Total Usage	Reservation Period	On-time Payment	Support Services Specification	Unit Price	Services Down Time	Penalty Rate
RightScale Cloud computing	-Editions: -Free -Standard -Premium -Corporate -Enterprise Solution Packs: Development and Test/Grid/Zend PHP HA/Social Gaming	Varies with services specifications, for example, S$1,255.10/month (initial fee S$5,020.40) for Premium Edition	[0, Upper Bound*]	1 month	Varies with services specifications, for example, S$627.55 + S$3,137.75 (initial fee for 5 customers) for Standard Edition	-Community -Bronze -Silver -Gold -Platinum	Included in solution charge	A 4 business hours to 3 business days response time guarantee for the support services	Information not available
Boomi	-Base -Professional -Enterprise	Varies with services specifications, for example, S$313.78 for extra connection for Standard Edition	[0, Upper Bound*]	1 month	Varies with services specifications, for example, S$690.31/month for Base edition	-Standard -Premier Response: 1 to 2 business days	-Standard Support is included in editions -Price for Premier depends on monthly subscription fee	A 99.99% monthly uptime guarantee for the services A 1 t 2 business days response time guarantee for the support services	A 10%-100% service credit if uptime not met
PiCloud	Computation	Varies with services specifications, for example, S$0.06/hour for c1-default instance	[0, Upper Bound*]	NA	NA	-Basic -Silver -Gold	-Basic: Free -Silver: $50 -Gold: Greater of $300 or 10% of monthly usage	A 99.9% annually uptime guarantee for the services	A 10% service credit if uptime not met

Appendix B. Proofs for Lemmas and Propositions

Proof of Lemma 1

It can be proved by comparing the expected utility a client gets from using fixed-price reserved services with unlimited resource capacity and that from using the same services with limited resource capacity.

Proof of Reserved Pricing Strategy Proposition (P1)

Vendor will maximize its profit, given by $\pi_{Reserved} = (1 - F(\lambda^*_{Reserved}))\cdot P_{Reserved}$. Plug in $F(\lambda^*_{Reserved}) = \lambda^*_{Reserved} / \Lambda$, and $\lambda^*_{Reserved} = P_{Reserved} / \bar{v}$, it can be solved that optimal price of reserved services contract is $P^*_{Reserved} = \bar{v}\Lambda / 2$.

Proof of Spot Pricing Strategy Proposition (P2)

Consider the bounded constraint $Prob_L \geq \frac{\gamma\, v_L}{(1+\gamma)\, v_L - P_L}$. Vendor's profit from spot-price services is $\pi_{Spot} = (P_L Prob_L^2 + P_H Prob_H) \cdot \frac{\Lambda}{2}$, which is no smaller than $\left(P_H - \frac{P_H^2}{4P_L}\right) \cdot \frac{\Lambda}{2}$. Rewriting the profit function we get $\pi_{Spot} = (P_L Prob_L^2 - P_H Prob_L + P_H) \cdot \frac{\Lambda}{2}$. Solve it we get vendor's optimal choice of $Prob_L$.

Proof of Corollary 1

The condition $P_H \geq P_L\left(1 + \frac{\gamma\, v_L}{(1+\gamma)\, v_L - P_L}\right)$ is equivalent to $P_H\big((1+\gamma)\, v_L - P_L\big) \geq P_L\big((1+2\gamma)\, v_L - P_L\big)$, which is equivalent to $P_H\big((1+\gamma)\, v_L - P_L\big)(v_L - P_L) \geq P_L\big((1+2\gamma)\, v_L - P_L\big)(v_L - P_L)$. Expanding the inequality we get $v_L^2 P_H(1+\gamma) - v_L(2+\gamma)P_L P_H + P_H P_L^2 \geq v_L^2 P_L(2\gamma+1) - v_L(2+2\gamma)P_L^2 + P_L^3$. Expanding the term $\left(P_L Prob_L^{*2} + P_H(1 - Prob_L^*)\right)\frac{\Lambda}{2}$ yields $\frac{v_L^2\left(P_L\gamma^2 + P_H(1+\gamma)\right) - v_L(2+\gamma)P_L P_H + P_H P_L^2}{\left((1+\gamma)\, v_L - P_L\right)^2} \cdot \frac{\Lambda}{2}$, and

plugging this term in to the condition just derived, further give

$$\frac{v_L^2\left(P_L\gamma^2+P_H(1+\gamma)\right)-v_L(2+\gamma)\,P_LP_H+P_HP_L^2}{\left((1+\gamma)\,v_L-P_L\right)^2}\cdot\frac{\Lambda}{2}\geq\frac{v_L^2\left(P_L\gamma^2+P_L(2\gamma+1)\right)-v_L(2+2\gamma)P_L^2+P_L^3}{\left((1+\gamma)\,v_L-P_L\right)^2}\cdot\frac{\Lambda}{2}=\frac{P_L\Lambda}{2}.$$

Therefore the profit in Condition (1) is greater than or equal to that in Condition (2).

Proof of Hybrid Pricing Strategy Proposition (P3)

The vendor's profit with a hybrid pricing strategy is $\pi_{Hybrid} = \left(1-F\left(\lambda^*_{Hybrid}\right)\right)$ $P_{Reserved} + F\left(\lambda^*_{Hybrid}\right)\cdot E\left(\lambda|\lambda<\lambda^*_{Hybrid}\right)\bar{P}_{Spot} = \left(1-\frac{2\tilde{v}-\bar{P}_{Spot}}{2\tilde{v}^2\Lambda}P_{Reserved}\right)P_{Reserved}$. Taking $\frac{2\tilde{v}-\bar{P}_{Spot}}{2\tilde{v}^2\Lambda}$ out of the equation yields π_{Hybrid} = $\frac{2\tilde{v}-\bar{P}_{Spot}}{2\tilde{v}^2\Lambda}\left(\frac{2\tilde{v}^2\Lambda}{2\tilde{v}-\bar{P}_{Spot}}-P_{Reserved}\right)P_{Reserved}$. This is a linear transformation of the same problem the vendor faces when it maximizes profit for fixed-price reserved cloud computing services. (See Proof of the Reserved Pricing Strategy Proposition.)

Proof of Impact of Spot-Price On-Demand Services Proposition (P4)

Subtracting π^*_{Hybrid} by $\pi^*_{Reserved}$ gives $\Delta\pi=\frac{\Lambda}{2}\left(\frac{\tilde{v}^2}{2\tilde{v}-\bar{P}_{Spot}}-\frac{\bar{v}}{2}\right)$. The term $(1+\gamma)Prob_LProb_H$ is bounded in $\left(0,\frac{1+\gamma}{4}\right)$. So, the profit difference $\Delta\pi$ will be positive if the condition $\frac{\bar{P}_S}{\bar{v}}>\frac{4(1+\gamma)Prob_LProb_H(1-(1+\gamma)Prob_LProb_H)}{4(1+\gamma)Prob_LProb_H-1+\sqrt{8(1+\gamma)Prob_LProb_H+1}}$ holds. The right hand side is bounded in $\left(\frac{(1+\gamma)(3-\gamma)}{4\left(\gamma+\sqrt{2\gamma+3}\right)},\frac{1}{2}\right)$. Since $\frac{(1+\gamma)(3-\gamma)}{4\left(\gamma+\sqrt{2\gamma+3}\right)}$ is in $\left(\frac{\sqrt{5}-1}{4},\frac{\sqrt{3}}{4}\right)$, as long as $\frac{\bar{P}_{Spot}}{\bar{v}}>\frac{1}{2}$, the profit difference $\Delta\pi$ will be positive. That means $\frac{\bar{P}_S}{\bar{v}}>\frac{1}{2}$ is a sufficient condition under which hybrid pricing improves the vendor's profit.

Proof of Impact of Consumer Surplus and Social Welfare Proposition (P5)

Consumer surplus in the hybrid pricing case is $CS_{Hybrid} = \frac{\bar{v}\Lambda}{2}\left(1 - \frac{\tilde{v}^2(3\tilde{v}-2\bar{P}_{Spot})}{\bar{v}(2\tilde{v}-\bar{P}_{Spot})^2}\right)$.

Let ΔCS be the difference between CS_{Hybrid} and $CS_{Reserved}$ (CS_{Hybrid} - $CS_{Reserved}$). Then $\Delta CS = \frac{\bar{v}\Lambda}{2}\left(\frac{3}{4} - \frac{\tilde{v}^2(3\tilde{v}-2\bar{P}_{Spot})}{\bar{v}(2\tilde{v}-\bar{P}_{Spot})^2}\right)$. The term $\frac{\tilde{v}^2(3\tilde{v}-2\bar{P}_{Spot})}{\bar{v}(2\tilde{v}-\bar{P}_{Spot})^2}$ can be transformed to $(1 + \gamma)Prob_L Prob_H \left(\left(\frac{(1+\gamma)Prob_L Prob_H}{2(1+\gamma)Prob_L Prob_H + \frac{\bar{P}_{Spot}}{\bar{v}}} - \frac{1}{2}\right)^2 + \frac{3}{4}\right) + \frac{\bar{P}_{Spot}}{\bar{v}}$, which is smaller than $(1 + \gamma)Prob_L Prob_H + \frac{\bar{P}_{Spot}}{\bar{v}}$. Since the first term $(1 + \gamma)Prob_L Prob_H$ is in $\left(0, \frac{1+\gamma}{4}\right)$, by plugging in its maximal to Δ, then plugging in Δ to ΔCS, I conclude that when $\gamma < (1/2 - 2\bar{P}_{Spot}/\bar{v})$, using hybrid pricing will increase consumer surplus.

Appendix C. Companions of the Experimental Study

Appendix C1. Analysis of the Influence of Job Duration on Spot-Price Services Performance

I began by analyzing the changes in spot price to obtain insights on services interruption risk. For spot-price data from Amazon for c1.xlarge instances in Western Europe from January 31, 2012 to January 31, 2013, I observed 1,195 price changes. The changes occurred any hour of the day. I simulated computing jobs with durations of up to 24 hours running on spot-price services to estimate the interruption risk associated with different job durations. Jobs of different types arrived randomly every hour during the days of the year, for a total of 8,760 jobs. (See Figures C1-1 and C1-2.)

I will formally describe the simulation in this paragraph. I did the simulations based on real prices from Amazon EC2 Spot market between January 31 8:00:00 AM,

2012 and January 31 8:00:00 AM, 2013. For each simulation, I fixed the duration of jobs, which can be 1 to 24 hours. I further assumed that in each clock hour only one job would arrive and immediately start running with spot-price services. The jobs would arrive at different time point in the clock hour though, controlled by a random number uniformly distributed in [0, 59] representing the minutes passed when a job arrived since the start of the clock hour. Therefore, each simulation contained 8760 jobs in total. The simulation agent was programmed to record for each job ran with spot-price services whether it completed in the end. The probability of services interruption for the job with fixed duration was then calculated as percentage of the 8760 job executions that were interrupted.

In addition, the simulation agent was programmed to also calculate and collect the percentage of completion of jobs that were interrupted, and the payment for each job ran with spot-price services. I in the end aggregated the results for all of the jobs and derived related statistics for the risk analysis. This process was repeated for the three experimental jobs with durations of 3, 5, and 10 hours.

The analysis results suggest that the longer the duration of a computing job, the higher the risk of services interruption. Though interruption can occur for job with any duration, overall job interruption is infrequent. For example, only 0.7% of all three-hour computing jobs running with spot-price services were interrupted due to price changes.

Figure C1-1. Spot-price changes at different clock times in the study period

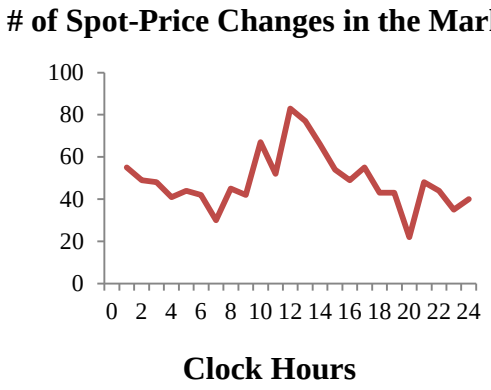
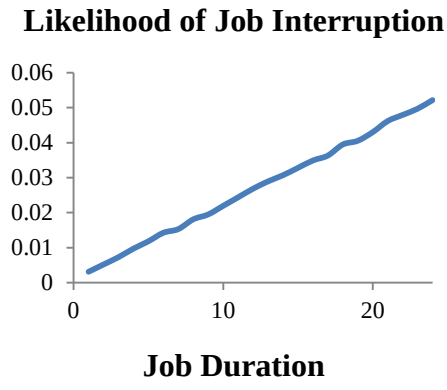


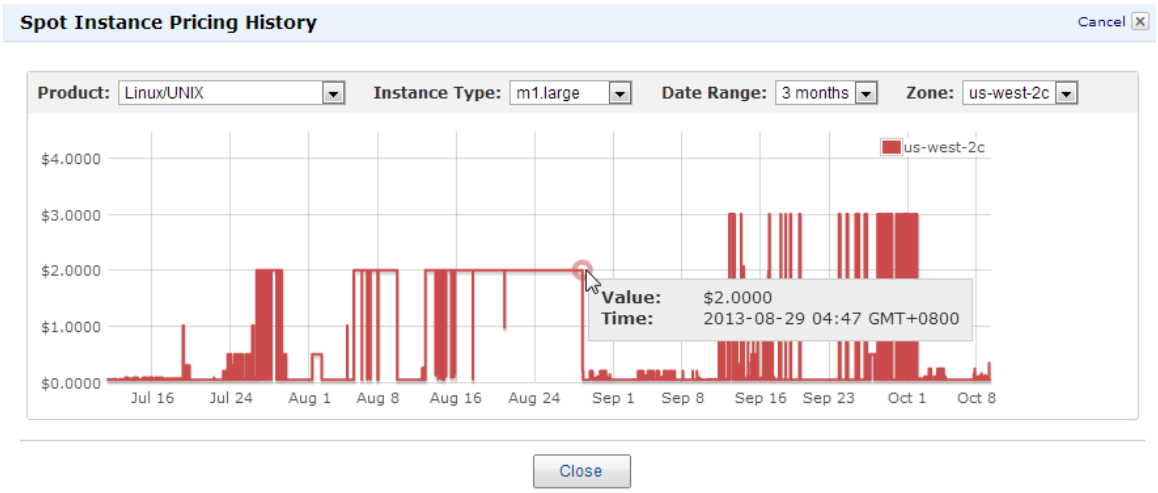
Figure C1-2. Interruption risk for computing jobs with 1 to 24 hour durations



My other ongoing field study of current spot-prices for cloud computing services suggests that job-risk analysis support is not available to clients from the vendors. Amazon.com provides real-time and historical prices for each type of computing instance it offers, only by region and computing capacity though. The historical price information gives clients a sense of the frequency of price changes but does not offer additional information to help clients to quantify the associated interruption risk for their jobs. (See Figure C1-3.)

The analysis suggests that: (1) due to unpredictable spot market price changes, every computing job run on spot-price services is subject to interruption risk; (2) computing jobs of longer duration bear a higher risk; and (3) clients do not have much risk analysis support from vendors.

Figure C1-3. Amazon.com's EC2 Management Console



Appendix C2. Risk Propensity Measurement (Adopted from Weber 2002)

For each of the following statements, please indicate the likelihood that you would engage in the described activity or behavior, if you were to find yourself in that situation. In addition, for each of the statements, please indicate how risky you perceive each situation to be. Provide a rating from 1 to 5 on the likelihood that you would engage in the described activity or behavior: very unlikely to neutral to very likely; how risky you perceive each situation to be: not at all – moderate – extremely risky --

1. Betting a day's income at the horse races.
2. Investing 10% of your annual income in a moderate growth mutual fund.
3. Betting a day of your annual income at a high-stakes poker game.
4. Investing 5% of your annual income in a very speculative stock.
5. Betting a day's income on the outcome of a sporting event.
6. Investing 5% of your annual income in a dependable and conservative stock.

7. Investing 10% of your annual income in a new business venture.

8. Gambling a week’s income at a casino.