Modeling and Optimizing Cloud Computing Service Prices

Ming-Wei (Caesar) Wu

ORCID: 0000-0002-2792-6466

Submitted in total fulfillment of the requirements of the degree of

Doctor of Philosophy

May 2019

School of Computing and Information Systems

THE UNIVERSITY OF MELBOURNE

Modeling and Optimizing Cloud Computing Service Prices

Ming-Wei (Caesar) Wu
Principal Advisor: Professor: Rao Kotagiri
Co-Advisor: Professor Rajkumar Buyya

Abstract

Modeling and optimizing cloud computing service prices are significant challenges facing many cloud practitioners and researchers in the field of cloud economics for either achieving competitive advantages or managing cloud resources effectively. Currently, the number of cloud pricing schemes offered by different cloud service providers (CSPs) is overwhelming. Many customers, especially business customers, find these pricing schemes puzzling and do not know how to analyze them to develop their business case so that they can transform their legacy IT infrastructure into a cloud platform. On the other hand, many new CSPs urgently need to know how to create and optimize the cloud price models so that they can effectively compete with their peers and serve their targeted customers well with limited resources. These are interdisciplinary challenges that involve cloud computing technologies, microeconomics, industrial organization, price theory, decision theory, market segmentation theory and value theory.

This thesis investigates these issues from both cloud customers and CSPs' perspectives. It provides cutting edge solutions to resolving the cloud price modeling and optimizing the problem. These proposed solutions are hedonic pricing for new cloud service features, cloud market segmentation, defining multiple customer utilities and cloud baseline pricing. This research advances state-of-the-art cloud economics and makes the main contributions as follows.

1. In light of the value theory, this study presents a comprehensive taxonomy and survey of the cloud pricing models proposed by many researchers during the last decade and identifies interdisciplinary challenges. It provides a unique way of classifying various models. It includes a short history of models and enabling technology-hypervisors.
2. By leveraging hedonic pricing for the new cloud service features, this work constructs different models that accord with customers' willingness to pay (W2P). To the best of my knowledge, it is the first time that cloud pricing is dependent on its service features that are differentiated by both intrinsic and extrinsic characteristics. Most significantly, this research unveiled the depreciation rate of cloud services, which is equivalent to Moore's Law.
3. Based on the market segmentation theory, this research generates a novel solution that combines both hierarchical clustering and time-series methods to extract the cloud customers' usage patterns from Google's public dataset and give a demand forecast from the local hosting firm's private dataset.
4. From the result of market segmentation and the demand forecast, six customer utility functions have been developed according to the Markov chain analysis, the queuing

theory (or M/M/s), and risk assessment for different business application workloads. The result of the utility function illustrates that cloud pricing should be built on market segmentation rather than on a unified market.

5. With a foundation of microeconomics, this research demonstrates a comprehensive fabric of a CSP's value-based cloud pricing strategy on how to generate different pricing models. In particular, it creates four models derived from cloud customer utility values and the CSP's cloud infrastructure costs. It also shows how to apply a genetic algorithm for identifying the optimal price point of each model for CSPs to maximize their profit.

Declaration

This is to certify that

1. The thesis comprises only my original work towards the Ph.D.,
2. Due acknowledgment has been made in the text to all other material used,
3. The thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies, and appendices.

Ming-Wei (Caesar) Wu, May 2019

Preface

This thesis research has been carried out in the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, School of Computing and Information Systems, Melbourne School of Engineering, The University of Melbourne. The main contributions of the thesis are discussed in Chapter 2 - 6 and are based on the following publications:

- **Caesar Wu**, Adel Nadjaran Toosi, Rajkumar Buyya, and Ramamohanarao Kotagiri, Hedonic Pricing of Cloud Computing Services, *IEEE Transactions on Cloud Computing, Cloud Computing*, (99), 1. DOI:10.1109/TCC.2018

- **Caesar Wu,** Rajkumar Buyya, and Ramamohanarao Kotagiri, Cloud Computing Market Segmentation, *Proceedings of the 13th International Conference on Software Technologies (ICSOFT 2018)*, ISBN: 978-989-758-320-9, Porto, Portugal, July 26-28, 2018

- **Caesar Wu,** Rajkumar Buyya, and Ramamohanarao Kotagiri, Cloud Pricing Models: Taxonomy, Survey and Interdisciplinary Challenges, *ACM Computing Survey*, Volume 52, No. 6, Article No. 108, Pages: 1-36, ISSN 0360-0300, ACM Press, New York, USA, October 2019

- **Caesar Wu** Rajkumar Buyya and Ramamohanarao Kotagiri, Value-based Cloud Price Modeling for Segmented Business Market, *Journal of Future Generation Computer Systems (FGCS)*, Volume 101, Pages: 502-523, ISSN: 0167-739X, Elsevier Press, Amsterdam, The Netherlands, December 2019.

- **Caesar Wu,** Rajkumar Buyya, and Ramamohanarao Kotagiri, Modeling Cloud Customers' Utility Functions, *Journal of Future Generation Computer Systems (FGCS)*, Volume 105, Pages: 737-753, ISSN: 0167-739X, Elsevier Press, Amsterdam, The Netherlands, April 2020.

- **Caesar Wu,** Rajkumar Buyya and Ramamohanarao Kotagiri. "Big Data Analytics = Machine Learning + Cloud Computing," *Big Data: Principles and Paradigms*, ISBN: 9780128053942, Waltham, MA Morgan Kaufmann, Elsevier, 2016. p. 3-37

Acknowledgments

Ph.D. is my childhood dream, which I always like to pursue during my lifetime. It is one of the significant milestones for the philosophical and scientific enlightenment during my lifelong voyage. Although it comes a bit later at my age and sometimes my research journey becomes very painful and exhausted, I am glad I could overcome numerous hurdles, including many physical barriers and see the light at the end of the tunnel. I cannot imagine I would be successful without my advisors' insight and advice. I also do not believe I would reach this milestone without countless help from people around me. With that, I would like to, first and foremost, thank my principal advisor, Professor Rao Kotagiri, who does not only offer me this precious opportunity but also patiently and vigorously guide me through this entire Ph.D. journey. I am so fortunate to have Professor Rao as my principal advisor.

I would also like to express my most profound appreciation to my co-advisor, Professor Rajkumar Buyya, who introduces me to Professor Rao. Without him, nothing would be possible. Without a joint effort of their supervision, I would still be searching in the dark. In addition to my advisors, I would like to take this opportunity to thank the chair of my advisory committee, Professor, Udaya Parampalli for his support to make progress of my research every step along the way. I would also like to take this opportunity to thank all the past and present members of CLOUDS laboratory at the University of Melbourne: Dr. Rodrigo Neves Calheiros, Dr. Amir Vahid Dastjerdi, Dr. Adel Nadjaran Toosi, Dr. Jungmin Son, Dr. Bowen Zhou, Dr. Xunyun Liu, Dr. Chenhao Qu, Dr. Maria Rodriguez, Dr. Safiollah Heidari, and etc. for them to share their experiences of their Ph.D. expedition.

I acknowledge the Australian Federal Government and the University of Melbourne granted the scholarships to me to pursue my childhood dream. This gift is beyond my imagination. I also express my sincere appreciation to Professor James Baily and Professor Justin Zobel for their lecture on how to implement a research project in the most effective and dialectic manner. Moreover, I would like to thank Dr. Natalie Karavarsamis for her excellent statistics lecture and make my rusty brain to think mathematically again after I left the academic world for almost 40 years. My sincere thank goes to the staff of the Melbourne School of Engineering including Rhonda Smithies, Julie Ireland, Emma Russo, Imbi Neeme, Sally Tape, Kate Hale, and all other Graduate research coordinators.

Lastly, I would like to thank all my family members, including my parents. I am deeply thankful to my son and my wife for their unconditional and daily support and encouragement. I am so fortunate that my son, Summa, can always offer his excellent opinion and valuable suggestions for all my research papers.

Ming-Wei (Caesar) Wu
Melbourne, Australia
May 2019

CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

Cloud price modeling and optimizing are the major challenges facing many practitioners and researchers in the field of cloud economics or Cloudonomics [1] due to cloud computing transformation and IT paradigm shift [3]. This implies that the number of new cloud service features is growing almost daily and the number of corresponding pricing schemes offered by different cloud service providers (CSPs) is overwhelming for different business applications. Consequently, it becomes a very complicated issue for many cloud customers, especially business customers, to make the right decision concerning the selection of the right pricing scheme for their business application needs during this transformation [39] [52]. Moreover, many incoming CSPs often want to know how to establish and optimize the cloud price models to maximize their profit while serving their targeted customers well and maintaining their cloud business to be competitive and sustainable. Yet many previous studies often focused on cloud pricing from a CSP's revenue and costs perspective (or internal rationality) and often paid less attention to the issue of the customers' utility value (or external rationalities).

Hence, this research will include both aspects of the cloud customer's utility values and CSP's profitability, and show how to construct various cloud price models and how to identify the optimal price point for each model according to both external and internal rationalities. In other words, it takes into account cloud market segmentation, market demand, customers' business applications and customer surplus value, as well as new cloud service features, the perishable asset effect and cloud infrastructure costs for CSP to maximize its cloud profits. As a result, this research will involve many disciplines, such as cloud computing technologies, microeconomics, industrial organization, operation research and value theory, to solve this complicated problem. Although the problem becomes very challenging, the strategy to tackle it is to divide it into two main problems and five manageable sub-problems:

1.) The first main problem is how to estimate a cloud price for new cloud service features: -

- How to estimate cloud pricing for new service features from a panel dataset, and
- How to estimate cloud pricing for new service features from a cross-sectional dataset.

2.) The second main problem is how to establish various price models for the cloud baseline or basic services:

- How to segment the cloud business market to understand cloud market demands;
- How to define the customers' utility values and functions, and
- How to develop various cloud pricing models for CSP's profit maximization.

The details of these sub-problems can be further elaborated as the first sub-problem is to deal with the issue of pricing new cloud service features or characteristics due to cloud technology innovation and new service offerings. The second sub-problem is to answer the question of cloud service pricing depreciation rate and lifecycle management. The third sub-problem is to tackle the challenge of capturing a greater market share, while the CSP can balance between uniform and personalized pricing for limited cloud resources. The fourth sub-problem is to define the cloud business customers' utility functions for their business application needs in the light of value co-creation. The fifth sub-problem is to show how to establish four cloud pricing models and how to identify the optimal price point for each pricing model.

Historically, various themes of computing pricing have been developed throughout the later 1990s, 2000s, and 2010s. Before the cloud computing era, Buyya et al. [51] attempted to establish economic models for grid computing. In 2009, they extended the idea of grid economics [302] to cloud price modeling with a market-orientation [52]. Yeo et al. [256] also provided an alternative idea of the pricing model, e.g., the automatic metered pricing model for utility computing services. Hande et al. [278] constructed a pricing model that is, with some constraints, of network accessing limitation. Xu et al. [80] presented a general pricing model that enables CSPs to charge their cloud customers according to the cloud infrastructure or resource needs. Moreover, Xu [56] developed a dynamic price model to combine both reserved and spot (or auction form) instance pricing schemes for a CSP to maximize its profits or revenue. Similarly, Toosi et al. [55] considered three different types of existing pricing models offered by Amazon Web Services (AWS) for CSP's revenue maximization. Ben-Yehuda et al. [61] used a dynamic algorithm to describe AWS' pricing models and speculated that the AWS pricing mechanism has a reserved price according to their observation of AWS spot instances' price history. Walker [77] [78], Greenberg [60] and Wu et al. [75] addressed the issue of cloud pricing model from a data center cost-based perspective

while El Kihal et al. [84] and Mitropoulou et al. [85] proposed a hedonic method to model the price and pricing index of cloud computing services.

The main themes of cloud services can be derived from the National Institute Standards and Technology (NIST) cloud definition of three service models: Infrastructure as a Service ((IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), four deployment models: Private cloud, community cloud, public cloud, and hybrid cloud and five characteristics: on-demand self-service, broad network access, resourcing pool, rapid elasticity, and measured service, which can be simply presented as "3S - 4D - 5C". The common sense of clouds computing is more like a new business model rather than a new technology ([297] [298] [299] [300] [301]) because it enables and redefines a new business relationship and a new value chain between cloud customers and CSPs in term of IT resource provisioning, delivery, deployment, consumption and management in cloud transformation. The stipulation of the cloud transformation leads to the central debate on the issue of how to make the right decision for cloud customers and how to make profit maximization for a CSP by establishing and optimizing various cloud price models. Weinman [1] recapitulated this topic in the new single word: "Cloudonomics" or how to apply an interdisciplinary approach for cloud computing service delivery.

Over the last few decades or so, many researchers have made excellent efforts for either cost-based pricing or resource-based pricing from a business to consumer (B2C) perspective, but these studies ([80] [86] [113] [114] [118] [119]) only provided a theoretical proof of the optimal pricing point for various pricing models and an explanation of existing models offered by various CSPs. Very little research has been done for customer values and market segmentation from the business to business (B2B) perspective. There is a significant gap in addressing how to create cloud pricing models and how to identify the optimal pricing point of these models based on the value-based pricing for business customer's application needs. It is urgently required to bridge this gap because the field of cloud computing has entered an "early majority" stage [15] and many business customers start to migrate their various workloads or business applications (e.g., mission-critical, e-Commerce, virtual desktop infrastructure (VDI), database backup) to the cloud infrastructure. They begin to consider a comprehensive cloud solution in terms of various cloud service features, end-user experiences and cloud ecosystems rather than a pure IT cost-cutting solution. From an evolutionary perspective of computing, IT pricing has been moving from "pay as you make" to "pay as you use" (billing method), and from a simple dashboard to auto-orchestration with OpenStack Application Programming Interface (API) due to cloud transformation. All these

trends are summarized in Figure 1—1 from the long-term evolution (or the ecosystem) of pricing model, service delivery and billing methods.

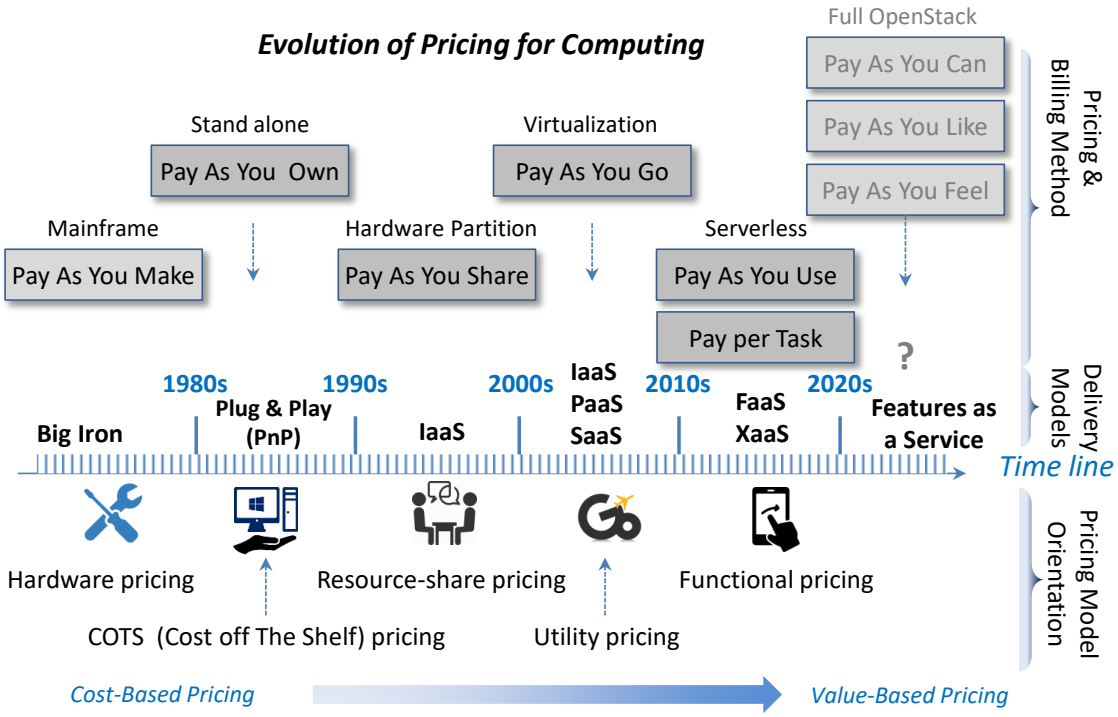


Figure 1—1 The evolutionary view of Cloud Computing Price Modelling [1]

On the basis of this evolution, the problem of definitions and consideration of previous theoretical investigations, this research will provide a total solution framework for the overall problem of pricing and optimization of cloud computing services. This total pricing solution consists of five sub-solutions (1.1, 1.2, 2.1, 2.2 and 2.3) that correspond to five issues (or sub-problems). The reason to divide the overall problem into two solution components with five smaller and manageable sub-solutions is that the overall problem is too challenging to be resolved as a whole. The first component is to model the cloud price based on the growing number of new characteristics (features) of cloud computing services. The new cloud service features may include security compliance, global data center footprints, OpenStack API, Burstable CPU, Failover, money-back guarantees, etc. The second component of the solution is to model the baseline service. It is not just to explain or prove the existing pricing model, but also to provide a

[1] OpenStack – This is an open source cloud management package that provides a common platform for controlling clouds of servers, storage, networks and even application resources. It is vendor-free management software.

practical solution for how to create various pricing models: for example, how to create the on-demand pricing model. The baseline service means the basic unit configuration of Infrastructure as a Service (IaaS), such as 1 CPU core, 4 GB RAM, 100 GB storage, 10 GBits network bandwidth in the specified data center location for one hour.

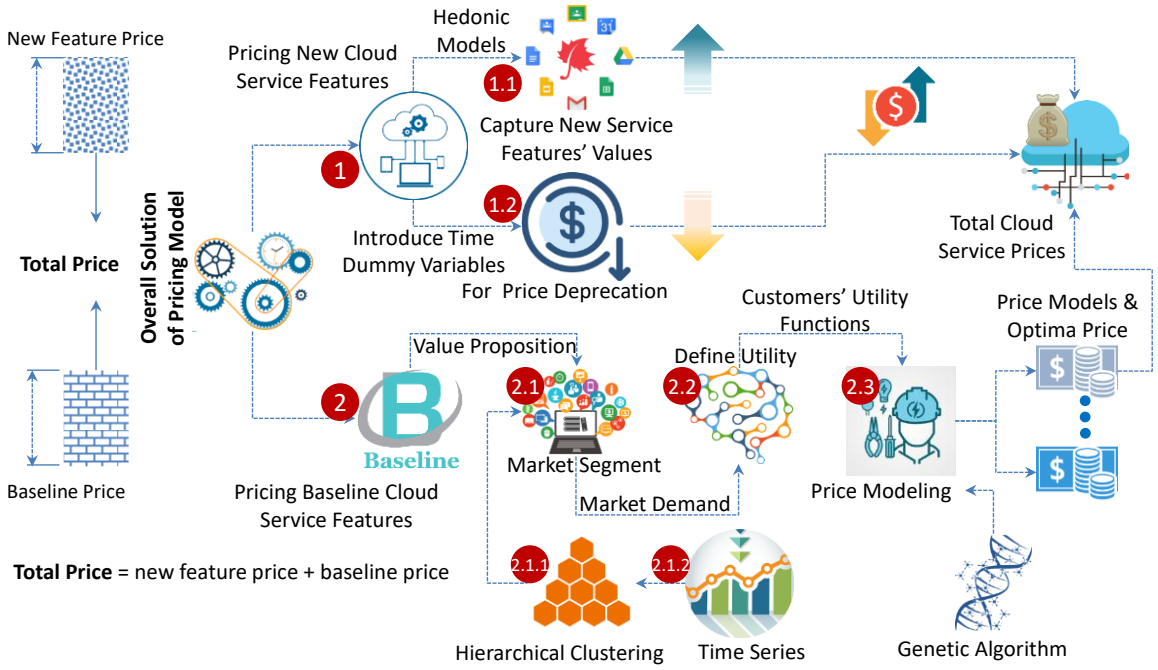


Figure 1—2 The Total Solution Framework of Cloud Service Pricing

Figure 1-2 illustrates the details of five sub-solutions: 1.1: adopting hedonic analysis to extract a new cloud service feature's price, 1.2: using time dummy to calculate a fixed effect for the price depreciation rate, 2.1: leveraging hierarchical clustering and time series to segment the cloud market, 2.2: defining cloud customers' utility functions based on a value co-creation principle, 2.3: building various cloud pricing models and identifying an optimal price for CSP's profit maximization by a genetic algorithm.

Overall, this research work proposes five practical sub-solutions for cloud price modeling that can be summarized from three aspects. 1.) This investigation establishes a comprehensive framework of cloud pricing from end to end for all cloud service features. 2.) It provides a novel solution that is not only for new cloud service characteristics pricing but also for cloud service lifecycle pricing. 3.) It establishes a fabric of value-based pricing strategy and demystifies the entire process of modeling and optimizing cloud prices for the baseline features.

In the context of the total solution of cloud pricing, the rest of this introduction chapter has five sections. The first section provides background information, which is to clarify some key concepts and terms for cloud price modeling and highlights my motivation behind this research work. The second section states the research problems and objects. The third section provides the evaluation method for various cloud price models that have been established. The fourth section gives a summary of the conclusions for this thesis. The fifth and final section draws a roadmap of how this thesis is organized and presented in detail.

1.1 Background

Suppose a hosting firm wants to extend its business to cloud computing services. The firm's executives or decision-makers decide to invest "x" amount of capital into cloud computing to grow its new cloud business. The subsequent question that the decision-makers wish to have answered is what the outlook of the cloud business revenue and profitability in terms of capital investment is likely to be. Intuitively, the fundamental problem is how to decide the cloud service pricing for various service features and how to estimate sales prices, market demand, and unit cost based on the defined business strategy (e.g. ,targeted customers, specified cloud service features, own service delivery expertise, technology strengths and a specified addressable market). From the microeconomics, the profit equation can be easily formalized as the relationship between the variables of sales price, market demand and unit cost (See Equations 1-1 and 1-2)

$$\pi[p] = \big[p - c_u[Q(p)]\big]Q(p) \tag{1-1}$$

where $\pi[p]$ is the profit, p is the sales price, $Q(p)$ is the market demand and $c_u[Q(p)]$ is the unit cost.

$$p = Q^{-1}(p) \tag{1-2}$$

While the definition of the equation appears to be very straightforward, an optimal solution to this equation is quite challenging because the relationship of p and $Q^{-1}(p)$ is intertwined. This study will provide a novel and innovative solution to this challenge by taking into consideration both internal and external rationalities for cloud customer's value proposition and for CSPs' profit maximization. The new solution of cloud pricing is driven by the idea of "value co-creation" ([270] [271] [272]), which emphasizes a partnership between cloud business customers and CSPs within the cloud market value chain. The origin of "value co-creation" can be traced back to the

motivation for this research, which is how to distribute the “good” value in the new cloud business value chain.

1.1.1 Motivations

During late 2008 and early 2009, I was involved in a business initiative and worked with a team to build cloud computing services capability for business customers. The initiative was the so-called “Project of Silver Lining” (PSL). The project was a natural extension of the existing hosting service business for many enterprise customers and government agents. The project aimed to target both existing and new customers' growth with a fixed amount of investment budget. In addition, this project also intended to gradually migrate many existing IT workloads from legacy infrastructure to a new cloud platform. It can also be considered as a cloud transformation or IT infrastructure lifecycle issue. When all issues were boiled down, the primary question was how to maintain the cloud business profitability and sustainability. It led to the issues of how to establish the cloud pricing models, how to identify the optimal price point for each pricing model, and how to capture a board spectrum of the values for various cloud B2B market segments.

If the pricing strategy is cost- and resource-based, the pricing solution is straightforward and relatively easily calculated where the markup price is a certain percentage of unit cost. However, this solution does not answer the question of customers' “willingness to pay” (W2P). It ignores external rationality. It could lead to a cloud price to be either overshot or undershot. T. Nagle et al. [10] demonstrated that cost-based pricing could become absurd. If it is market-based pricing, many new and innovative cloud features would not exist in the current market because there is no existing market environment to reflect supply and demand. The logic solution is “value-based” pricing because it is considered to be a better approach to price the services [36]. Nonetheless, it is very ambitious because value-based pricing is subjective, arbitrary and subtle. It is often hard to quantify and measure, especially for cloud services within the B2B market. Philosophically, the value-based pricing should determine the values contribution to the customer’s business from three aspects: “Good to have” (consolidate values); “Good to do” (build new values), and “Good to be” (grow future values) [294]. The puzzle of value-based pricing motivates me to investigate the cloud pricing further and move beyond just the cost-based and market-based price models.

1.1.2 Cloud Computing Service Price Modeling

If the value-based pricing for cloud service concerns both internal and external rationalities, the logical question is, what is the relationship between value-based, cost-based and market-based pricing? Based on the value theory, which is to study the nature of value evaluation, cloud price modeling can be classified into three basic strategies that have already been mentioned above: value-based (subjective) pricing (refer to Section 2.2.3.1), market-based (interactive) pricing (refer to Section 2.2.3.2) and, cost-based (objective) pricing (refer to Section 2.2.3.3). Mathematically, the relationship of the three pricing strategies can be derived from the Lerner Index and presented in the following Equation 1-3 and Figure 1—3.

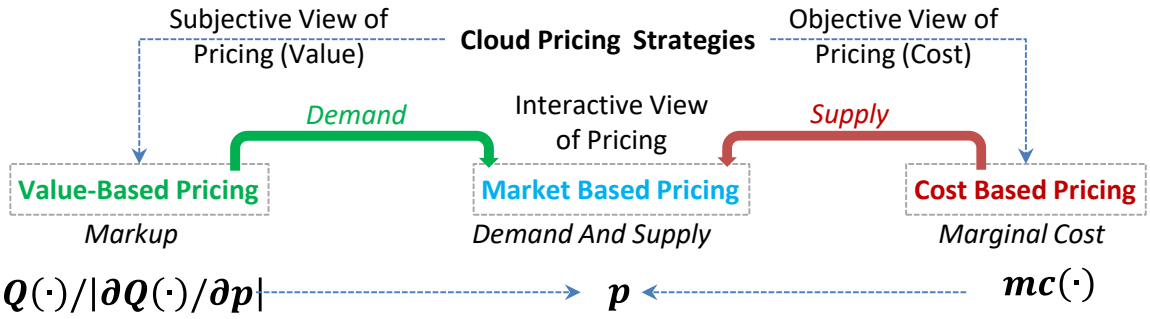


Figure 1—3 Three cloud pricing Strategies

$$p = mc(\cdot) + \frac{Q(\cdot)}{|\partial Q(\cdot)/\partial p|} \tag{1-3}$$

where $mc(\cdot)$ is the marginal cost or average cost, p is the market price and $Q(\cdot)$ is the market demand, $\frac{Q(\cdot)}{|\partial Q(\cdot)/\partial p|}$ represents the markup price. In this equation, the markup price is the primary factor that determines the value-based pricing because it is ultimately derived from the cloud customer's "willingness to pay" (W2P). This W2P is dependent on the value estimation for various cloud service features. What is the cloud service feature?

1.1.3 Cloud Service Characteristics

Cloud service features and characteristics are often used interchangeably. They are innovative attributes of cloud services for CSPs to gain competitive advantages in the cloud market place. For example, AWS has grown its new innovative cloud service feature almost daily since 2006. Up to the end of 2017, AWS has released a total of 1,430 new features [295], such as burstable CPU, Elastic Cache, IoT platform, Greengrass, Lex, and AWS Lambda FaaS, which is in contrast to a baseline cloud service. Often, a CSP marks up a price for baseline service (e.g., Infrastructure

as a Service or IaaS) that may include different cloud service features. Different CSPs may offer different baseline services. It is very challenging for a cloud customer to compare cloud prices for different baseline services that may include different cloud service features.

If we just compare the price of IaaS among 30 leading global CSPs, AWS's price is not the cheapest. AWS's price of its IaaS is slightly above the median price. But AWS can still maintain over 31% of its IaaS global market share and keep double-digit revenue growth year on year (YoY). This phenomenon, together with the pricing puzzle of PSL, inspires me to formulate the following research topic and objectives.

1.2 Research Problems and Objectives

The purpose of this thesis is to address the question of cloud price modeling and optimizing for a value proposition for both CSPs and cloud customers. The research topic can be defined as

How to predict, model and optimize the prices of both new and baseline cloud computing services based on the value-based pricing strategy for a CSP to maximize revenue or profit that allows the CSP to develop a competitive and sustainable cloud service business. This complicated problem can be solved by five proposed solutions, shown in Figure 1—2.

In order to implement this research, the following research objectives have been identified.

- Classify and survey the current state of the art of cloud computing price modeling systemically in order to fully understand the significant findings, controversies, and gaps of previous works.
- Propose a practical solution for cloud services pricing by consideration of its ecosystem, which has to cope with an ever-changing environment of cloud infrastructure. This investigation will apply a hedonic analysis to modeling for both cloud intrinsic and extrinsic features of cloud services based on the utilitarian theory, as shown in Figure 1—2. This study will implement three experiments: a.) Test intrinsic characteristics' impact on pricing by leveraging the AWS dataset; b.) Test time dummy's impact on pricing by leverage the AWS panel dataset of the past 10 years, and c.) Test extrinsic characteristics' impact on pricing by leveraging five leading CSPs' cross-sectional dataset in one year.

- Implement the hierarchical clustering and time series algorithms to segment and predict the market demand for the cloud B2B market based on Google's public dataset.
- Define and create different cloud customer utility functions according to the cloud market segment for cloud pricing modeling and optimization. This research will apply Markov Chain analysis, queueing theory and risk analyses for various cloud customers' utility functions
- Create various cloud pricing models and identify the optimal price point of each model for a CSP to achieve business revenue and profit maximization. This study proposes to establish four possible pricing models according to customers' surplus values and adopt the genetic algorithm to identify the optimal price point for each proposed price model.

1.3 Evaluation Methodologies

Regarding pricing new cloud service features, this study adopts the regression analysis forecasting method to predict cloud prices. The process to evaluate the predicted performance is first to create a robust hedonic function that consists of intrinsic, extrinsic and time dummy variables from both panel and cross-sectional datasets. The second step is to use the generated hedonic function to predict future price point of 2017 from a 2014 dataset. The accuracy of prediction results shows as high as up to 78.6%.

For the cloud market segmentation, this investigation employs the majority rule method to determine and evaluate the optimal number of cloud market segments. R's NbClust package has developed more than 30 methods or indices to evaluate the optimal number. This study used the NbClust package to evaluate the number of cloud market segments. Regarding the cloud market demand evaluation, the Time Series estimation is assessed by the Auto-Correction Function (ACF) to check the residuals movement based on the local hosting firm's dataset. The final result is also verified by the forecast data published by global leading technology research firms, such as Gartner, IDC, Forrester Research and ISG.

The methodology used to evaluate cloud customer utility function is to compare a CSP's profit margin, sales volume, unit cost and optimal price with the existing state of the art, such as simple linear and resource-based pricing based on the uniform market assumption. The experiment's

results show that the proposed multiple utility function solution can improve the CSP's profit margin by over 213% compared to current solutions.

On the assumption that a CSP offers four various pricing models in the cloud market, this study is also to evaluate the business performance in terms of that of CSP's profit margin, sales volume, optimal price and unit cost between cost-based and value-based pricing. The markup assumption for cost-based pricing is 100%, while the value-based pricing is dependent on customers' surplus values or utility functions. The performance results show that although the sales volume is the same, the profit margin is over 100% compared with resource-based pricing. In other words, the value-based pricing can achieve over 200% profit margin.

1.4 Thesis Contributions

The significant contributions of this work can be classified into five categories. The first is a survey and taxonomy of cloud pricing models. In contrast to previous surveys and taxonomy, this work is based on value theory to "carving (cloud pricing) the nature at its (economic) joints" [295]. Secondly, this work proposes a novel hedonic model to capture cloud extrinsic characteristics or innovative service features. The third category is to classify the cloud market into different segments by hierarchical clustering and time series method. The fourth one is to identify multiple cloud customers' utility functions along various market segments to lay out a framework of value-based pricing strategy. Lastly, this thesis demystifies and establishes the complete process of cloud price modeling through four different value-based models for CSP to maximize its cloud business profits.

The primary contributions of this thesis can be further articulated in the following details.

1. A taxonomy and literature survey of cloud models provides an overview of cloud pricing evolution of both practice and theory and highlights an outlook of cloud price modeling in the context of value theory with a multiple-discipline approach of economics, industrial organization, price theory, market theory and cloud computing technologies.

2. Formalization of a hypothesis that cloud services have both intrinsic and extrinsic characteristics based on utility theory. This then tested and validated a cloud pricing model according to both an AWS panel and five leading CSPs' cross-sectional datasets.

- A novel hedonic model is established that can differentiate three variables, namely, intrinsic, extrinsic and time dummy.
- With the novel hedonic model, a CSP can identify how much a customer is willing to pay. The experimental results showed that the extrinsic value is about 43% above the baseline service.
- The hedonic model has also unveiled the cloud service average annual growth rate (AAGR) at - 20% (or depreciation rate) between 2008 and 2017, which is at a much slower pace than with Moore's law.
- The research work of the hedonic pricing model illustrates that CSPs should not compete on price but rather price based on innovative service features.
- The model provided a less biased pricing model for many cloud decision-makers to develop their investment strategies in the cloud business.

3. According to the classical theory of market segmentation, a novel solution is proposed to segment the B2B cloud market. It consists of both supervised and unsupervised learning algorithms that can precisely quantify the cloud market segment that lays out the foundation for CSPs to construct various cloud pricing models.
   - Demonstrate how to use hierarchical clustering algorithms to identify the optimal number of cloud market segments by leveraging Google's public cloud dataset as the input to extract cloud customer usage patterns.
   - Show how to assess the clustering tendency that contains the cloud usage pattern.
   - How to determine the number of market segments and the proportion of each segment and how to validate the cloud market segments statistically.
   - How to adopt the Time Serial method to predict local cloud market demand.
   - Illustrate how to combine the predictable cloud market demand with the segmented market.

4. Define various utility functions in terms of multiple cloud market segments. It is built upon the various utilities by a Markov Chain analysis, queue theory and risk assessment.
   - Use cloud customers' revenue and profit as a measurement quantity for cloud customers' utility level, which directly reflects cloud customers' fundamental value proposition.

- Adopt a Markov chain analysis to calculate the cloud customer's business value in terms of SLA measurement and requirements for a cloud business customer.
- Leverage the queuing theory, the customers' utility function to reflect the minimum response time in terms of cloud customers' revenue impact for e-Commerce business application, such as an online checkout system.
- By microeconomics, the utility functions of business customers are defined by a risk assessment, in which the utility function is dependent on the relationship between a risk factor and the cloud resources.

5. According to the combined result of cloud market segmentation and cloud customer utility functions, four value-based cloud price models, namely, on-demand, bulk, reserved, and bulk + reserved, are established and then optimized. In contrast to previous cost- or resource-based pricing, the pricing models include both internal and external rationalities. With the proposed genetic algorithm (GA), CSP's revenue and profit can be maximized for each pricing model.
   - Formalize four value-based cloud price models as a business solution that allows CSPs to maximize the profit and revenue for the B2B cloud market, which enables CSPs to develop a business partnership with their business customers to achieve the value co-creation or a business partnership.
   - By developing various cloud pricing models, it allows CSPs to capture a broader spectrum of cloud market share and customers' surplus values.
   - Illustrate how to optimize these cloud price models using GA.
   - Cloud price models are dependent on both internal (cloud infrastructure costs) and external (customer utility functions and market segments) rationality.
   - Show that the bulk-selling + reserved pricing model can achieve the best business revenue and profit margins in comparison with other pricing models.

1.5 Thesis Organization

The structure of the thesis is organized in seven chapters and is shown in Figure 1—4. These are based on the number of publications published during my Ph.D. candidature. The overall

research topic is seen as having two themes: one is devoted to new cloud service characteristic pricing and the other addresses baseline service pricing.

- **Chapter 2** develops taxonomy and a survey of cloud pricing models. The content of this chapter is derived from the following paper:
  - Caesar Wu, Rajkumar Buyya and Kotagiri Ramamohanarao, "Cloud Pricing Models: Taxonomy, Survey and Interdisciplinary Challenges," *ACM Computing Surveys, Volume 52, No. 6, Article No. 108, Pages: 1-36, ISSN 0360-0300, ACM Press, New York, USA, October 2019.*

- **Chapter 3** provides hedonic price modeling for cloud computing services. It mainly focuses on new cloud service features or characteristics. The content of this chapter is underpinned by the following paper:
  - Caesar Wu, Adel N. Toosi, Rajkumar Buyya and Kotagiri Ramamohanarao, "Hedonic Pricing of Cloud Computing Services" *IEEE Transactions on Cloud Computing, Cloud Computing, IEEE Transactions on, IEEE Trans. Cloud Computing, no. 99, p. 1-15*.

- **Chapter 4** presents a novel solution to the segment cloud computing market segment, which is based on the theory of market segmentation. This chapter is built upon the following paper:
  - Caesar Wu, Rajkumar Buyya, and Kotagiri Ramamohanarao, "Cloud Computing Market Segmentation" *ICSOFT proceeding, 2018, p. 888-897*

- **Chapter 5** exhibits how to model cloud customers' utility functions based on various applications in different market segments. This chapter is determined by the following paper:
  - Caesar Wu, Rajkumar Buyya, and Kotagiri Ramamohanarao, "Modeling Cloud Customers' Utility Functions," *Journal of Future Generation Computer Systems (FGCS) Volume 105, Pages: 737-753, ISSN: 0167-739X, Elsevier Press, Amsterdam, The Netherlands, April 2020.*
  - Caesar Wu, Rajkumar Buyya and Ramamohanarao Kotagiri. "Big Data Analytics = Machine Learning + Cloud Computing," *Big Data: Principles and Paradigms,*

ISBN: 9780128053942, Waltham, MA Morgan Kaufmann, Elsevier, 2016. p. 3-37

- **Chapter 6** proposes four value-based cloud price models based on cloud customers' utility functions and cloud market segmentation. This chapter is developed in the following paper:
  - Caesar Wu, Rajkumar Buyya, and Kotagiri Ramamohanarao, "Value-based Cloud Price Modeling for Segmented Business to Business Market," *Journal of Future Generation Computer Systems (FGCS) Volume 101, Pages: 502-523, ISSN:0167-739X, Elsevier Press, Amsterdam, The Netherlands, December 2019.*

- **Chapter 7** provides the summary information about this thesis together with an analysis and discussion on future direction. This chapter is again derived from the following paper:
  - Caesar Wu, Rajkumar Buyya, and Kotagiri Ramamohanarao, "Cloud Pricing Models: Taxonomy, Survey and Interdisciplinary Challenges," *ACM Computing Surveys, Volume 52, No. 6, Article No. 108, Pages: 1-36, ISSN 0360-0300, ACM Press, New York, USA, October 2019.*

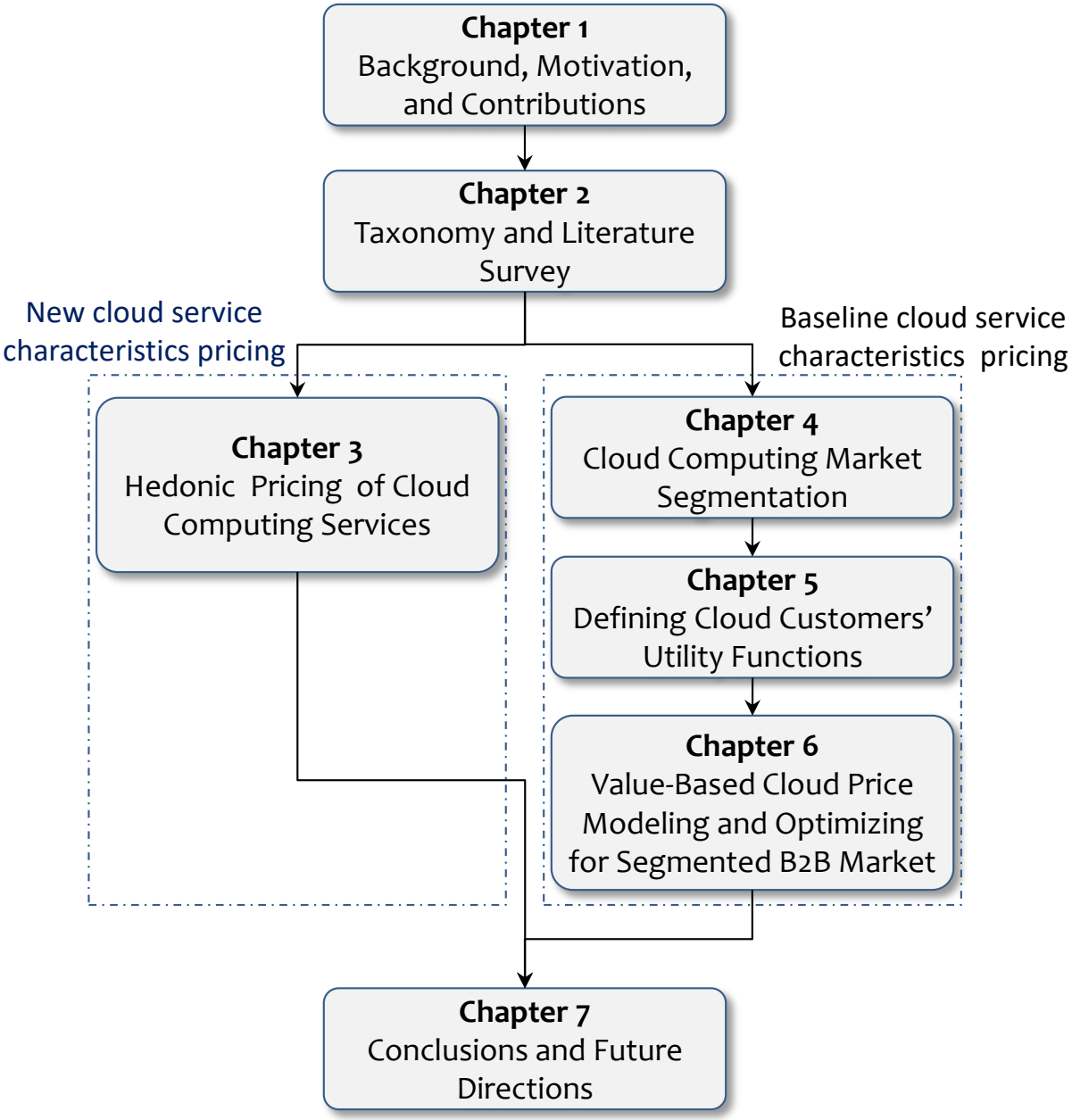


Figure 1—4 Thesis organization

Chapter 2

Taxonomy and Literature Survey

This chapter provides a systematic overview of cloud pricing in an interdisciplinary approach. It examines many historical cases of pricing in practice and tracks down multiple roots of modeling in research. The purpose of this overview is to help both cloud service providers (CSPs) and cloud customers to capture the essence of cloud pricing when they need to make a critical decision either to achieve competitive advantages or to manage cloud resources effectively. Currently, the number of available pricing schemes is overwhelming. It is an intricate issue to understand these schemes clearly due to involving several domains of knowledge, such as cloud technologies, microeconomics, operations research, and value theory. Some earlier studies have introduced the cloud pricing models unsystematically. Their approaches inevitably lead to much confusion for many cloud decision-makers. Consequently, this chapter presents a comprehensive taxonomy of cloud pricing models, which is driven by a framework of three fundamental pricing strategies that are built on 9 cloud pricing categories. These categories can be further mapped onto a total of 60 pricing models. Many of pricing models have been already adopted by CSPs. Others have been widespread across other industries. This study gives descriptions of these model categories and highlights both advantages and disadvantages. Moreover, this chapter offers an extensive survey of many cloud pricing models that were proposed by many researchers during the last decade.

2.1 Introduction

Cloud Computing transformation is now taking a momentum [12] [13]. It has entered the stage known as "early majority" of the cloud technology adoption life cycle, in which cloud computing

This chapter is derived from:

- **Caesar Wu**, Rajkumar Buyya, and Kotagiri Ramamohanarao, "Cloud Pricing Models: Taxonomy, Survey and Interdisciplinary Challenges," ACM Computing Surveys, Volume 52, No. 6, Article No. 108, Pages: 1-36, ISSN 0360-0300, ACM Press, New York, USA, October 2019

has become a mainstream market of IT infrastructure [15]. According to Wikibon [14], the Compound Annual Growth Rate (CAGR) of true private cloud (a hyper-converged cloud solution) alone will grow 29.2% from 2017 to 2027 while IaaS will grow 15.2% during the same period. However, one critical issue has remained unclear, which is how to understand a variety of cloud pricing models that are offered by different Cloud Service Providers (CSPs). Yet, the number of pricing schemes in the current cloud market is overwhelming. The aim of this chapter is to provide a systematic overview of many pricing models for both CSPs and cloud customers [3] so that CSPs can achieve its cloud business competitiveness and sustainability while cloud customers can make the right decisions during this cloud transformation.

Recently, many CSPs or cloud computing advocators claim that cloud computing is cheaper computing due to its Total Cost of Ownership (TCO) [16] [17]. However, Weinman [1] argued that "Cloud Computing is not cheap computing." Martens et al. [2] echoed this view, and they have noticed that many cloud cost (price) conclusions lack a systematic approach for a cost estimation regarding various models. Many favored claims are often dependent on an ad-hoc processing approach without consideration of some indirect and hidden variables.

As a result, Buyya et al. [3] [4] suggested that the topic of cloud computing pricing should be considered in an interdisciplinary way, which should be studied under the scope of multiple disciplines including cloud technologies, price theory, microeconomics, operations research, and value theory. Similarly, Kash and Key [306] also indicated "current cloud pricing schemes are fairly simple." "Multidimensional scheduling and pricing offer greater potential for increasing both customer satisfaction and (CSP)'s revenue" with a growing number of new cloud service features. According to [5] [6], no single discipline can provide a satisfying solution for cloud pricing. An isolation approach of cloud pricing could increase the difficulty for decision-makers to comprehend the benefits and risks of cloud services as well as a price to be paid. One of the examples is how to understand Amazon Web Services (AWS) spot instance or spot block (up to 6-hour service duration time) pricing. It can be considered as dynamic-based pricing [4] [37] [38]

[3] Cloud business customers have their own business, such search engine optimization (SEO), storage backup, virus scanning and etc., run on the cloud infrastructure to serve other customers. They are not end users. From a cloud customer's perspective, CSP's cloud price is equivalent to its cost.

[4] The dynamic pricing model means the price is a function of many variables, such as time, season, customer demand, etc. Many firms adopt this price to manage their yield for their limited capacity or resources. It has been widely applied in many service and utility industries such as airline, hotel, electric and gas utilities.

because of the nature of fluctuation [97]. On the other hand, it can also be regarded as auction-based or even cost-based pricing because of its multiple characteristics [98] [99]. Therefore, this chapter argues the cloud pricing issue must be investigated by an interdisciplinary approach from a value proposition perspective.

Although this survey draws multiple disciplines for the pricing issue, it mainly focuses on four knowledge domains: the Cloud pricing model is the focal point. Microeconomics is the theoretical tool to understand the cloud price that is influenced by supply and demand in the cloud market place. Value theory is the measurement of a customer's value proposition, which is defined by a CSM. Operation research is a method to help cloud decision-makers to make better decisions for any given cloud price during cloud transformation. Cloud technologies allow CSP to establish various new cloud pricing models to capture maximum customer surplus values from multiple cloud market segments. Throughout this taxonomy and survey, we will examine both the pros and cons of different cloud pricing strategies [5] and models [6] regarding a fundamental question of value [100]. We will also investigate pricing models to reflect the subjective experiences of many cloud business customers. These subjective experiences are often measured by Cloud Service Metrics (CSM) [103] [104], such as acquisition, retention, and efficiency from a business customer's perspective.

Overall, this work derives from three strategies of cloud pricing through both subjective (values) and objective (fact) views. Value-based pricing is demand-driven, and cost-based pricing is supply-driven. Moreover, market-based pricing can be seen as the result of an equilibrium of both supply and demand in a cloud market. Based on these basic strategies, we can define a hierarchical pricing framework that is illustrated in Figure 2—1. Each layer of the framework is driven by its goal. At the top, the pricing is driven by the principle of value [100]. The next layer down is derived from three pricing strategies, which are to pursue a long-term goal of a business. The layer further down is drawn from pricing tactical [7], which is oriented by short-term objects. The aim of tactical pricing is how to translate a pricing strategy to tactical objects. Finally, the bottom

[5] Strategy is how does a decision maker deal with or solve the given business problem for a long term or overall goal.
[6] Model is a representation of strategy. It can help us to visualize and access the relationship of the various objects. It is a simplified or abstracted description of reality, especially a mathematical one, for us to predict the future.
[7] Tactic is similar as a strategy, which is a plan to achieve a specified aim. However, the aim of a tactic is to gain immediate or short-term benefits rather than long term one. It is possible to win a game tactically but lose it strategically. Many tactics can support an overall strategy.

layer of cloud pricing consists of 60 individual models, who are detail-oriented. It explains the details of implementing a pricing strategy. The meaning of this framework implies if a strategy is cost-based, the final price "*p*" is determined by a cost that is driven by internal rationality. In contrast, if a strategy is value-based, "*p*" is dependent on cloud customers' utility value, which is mainly determined by external rationality. If a CSP decides to adopt market-based pricing, "*p*" is a result of the market equilibrium of supply and demand. The essence of this hierarchical framework can reflect the microeconomics [9] in term of price theory.

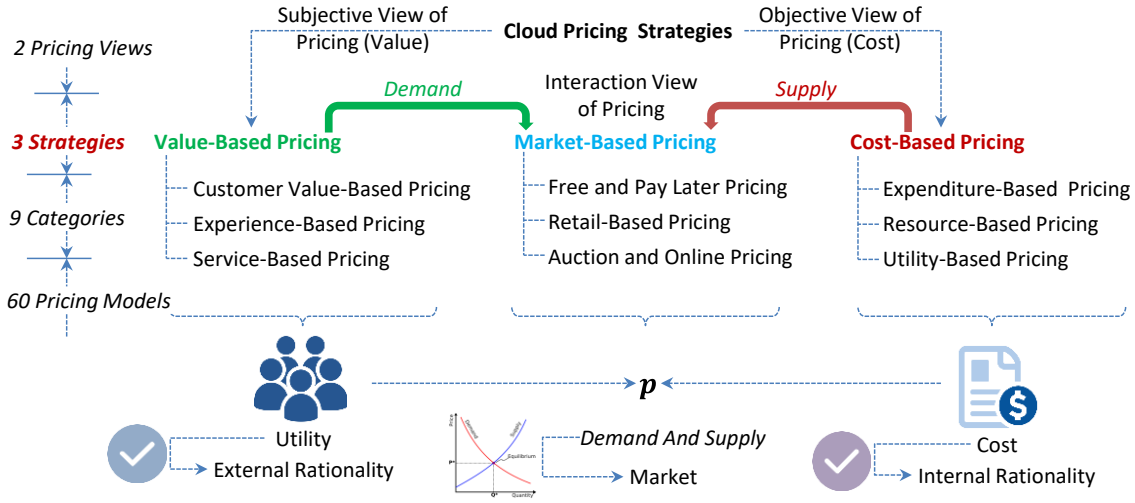


Figure 2—1 Big Picture of Multiple Disciplines of Cloud Pricing Models

According to this framework, we can find many earlier works mainly focused on both cost-based and market-based pricing and paid less attention to value-based pricing. Therefore, this study will not only include all three pricing strategies into consideration but pay special attention to value-based pricing. The primary reason is if a CSP knows all the pricing components (facts) of cloud (such as cloud service cost, markup ratio, market share and target rate of return, etc.) [7] [8] objectively, a cloud price still can't be determined because a decision-maker does not know how to handle these facts, which item (fact) is more important than the others, why and when it is much more important than others. These questions are the question of value [19]. If someone would insist on derivate from the fact to value alone, it becomes a naturalistic fallacy [20], which T. Nagle et al. [10] demonstrated that this kind of pricing strategy would become absurd. In order to avoid this logic fallacy, this work will provide a comprehensive framework by considering both CSP's cost and customers' value proposition for cloud pricing. As a result, this chapter has made the number of contributions listed as follows:

- It establishes the unique framework of classifying various cloud pricing models
- It categorizes 60 pricing models into three pricing strategies and nine pricing categories. Many models have not been considered by CSPs yet, but they have been widely adopted by other service industries, such as airline, travel, hotel, recreation, healthcare, telcos, and retail sectors. The purpose of exploring these potential models is to help many CSPs to compete on pricing, not on a price.
- It reviews most of the recently proposed models in considerable depth regarding their contributions and gaps plus their business application. Moreover, our work also highlights characteristics of pricing models offered by leading CSPs, which they often leverage their business strength to build their models.
- It provides the key for many cloud decision-makers to comprehend various cloud pricing models easily.

The rest of the chapter is organized as follows: Section 2.2 reviews the history of cloud pricing from a practical perspective. It includes cloud service launch times and virtualization technologies that underpin various cloud prices and cloud business. The aim of having this historical overview is to understand the multiple roots of cloud pricing models proposed by many researchers during the last decade. This work then outlines a relationship three pricing strategies based on value theory. Section 2.3 establishes the taxonomy of cloud pricing models. Section 2.4 provides a detailed survey of selected papers that were published from 2008 to the present. Finally, this chapter compares each pricing model with other models for its methodology and theoretical roots (All acronyms in this Chapter are listed in Table 2—1)

Table 2—1 Acronym Used In This Chapter

Acronyms	Definition	Acronyms	Definition	Acronyms	Definition
AWS	Amazon Web Service	DevOps	Development and Operation	NoOps	No Operation System
API	Application Programming Interface	EC2	Elastic Compute Cloud	Opex	Operation Expenditure
B2B	Business to Business	E2E	End to End	QoS	Quality of Service
B2C	Business to Customer	GCP	Google Cloud Platform	PAYG	Pay As You Go
CAGR	Compound Annual Growth Rate	KVM	Kernel-based Virtual Machine	S3	Simple Storage Service
Capex	Capital Expenditure	HARA	Hyperbolic Absolute Risk Aversion	SCADA	Supervisory Control& Data Acquisition
CBA	Cost-Benefit Analysis	DaaS	Database as a Service	SEO	Search Engine Optimization
CoD	Code on Demand	IaaS	Infrastructure as a Service	SLA	Service Level Agreement
COTS	Cost off The Shelf	PaaS	Platform as a Service	SME	Small Medium Enterprise
CPI	Customer Price Index	SaaS	Software as a Service	SOA	Service-Oriented Architecture
CRM	Customer Relationship Management	FaaS	Function as a Service	TCO	Total Cost of Ownership
CRRA	Constant Risk Aversion	XaaS	Anything as a Service	TPU	Tensorflow Process Units
CSM	Cloud Service Metrics	NPV	Net Present Value	VM	Virtual Machine
CSP	Cloud Service Provider	NPC	Net Present Capacity	W2P	Willingness to Pay

2.2 History of Cloud Pricing Models

2.2.1 Cloud Pricing Models In Practice

The first cloud pricing model can approximately be traced back to Salesforce.com's Russian doll model [8] [101], which are similar to optimal feature pricing (one of the retail-based pricing models). We can also consider it as per-user-based pricing for Software as a Service (SaaS). It is dependent on a value proposition. Salesforce.com's pricing model is a contrast to Siebel's distributed or perpetual licensing model. Back in 2000, the average price of Siebel's Customer Relationship Management (CRM) software would be around $10,000 per license plus additional $5,000 ongoing costs for a patch, regular upgrades, bugs fixes, maintenance, backup, and help

[8] Russian Doll or Matryoshka Doll pricing model is a type of marketing strategy to bundle different product features into one nested deal, like a Russian Doll.

desk support. Consequently, it is beyond many small and medium enterprise (SME) customers reach because they could not afford to allocate a significant amount of IT budget or Capital expenditure (Capex) upfront. This issue led to an opportunity for Marc Benioff (one of the founders of Saleforce.com) to offer a subscription-based pricing model for SaaS [21].

The cloud technology that underpins per-user-based pricing is known as software multi-tenancy. The idea of multitenancy is an analogy to drawing from an apartment building where the tenants can share the cost, such as public facility, body-corporation, security, etc., but still have their private space. By the same principle, Microsoft Hotmail or Google's Gmail also offers the email service, which every user (or tenant) can enjoy the email service via a web browser without any stress of installation and configuration of the mail software by themselves. _Figure 2—2 summarizes a timeline of different pricing models that were adopted by some leading CSPs along with cloud technologies development.

Following the similar concept of sharing, AWS, one of the global leading IaaS providers adopted the "on-demand" pricing model for its Simple Storage Services (S3) that was launched in Mar 2006 and offered Elastic Compute Cloud (EC2) in Aug 2006 for its public cloud. The enabling technology of both S3 and EC2 was Xen hypervisor, which Citrix Systems released the initial version in Oct 2003. Later in 2009, AWS launched its spot instance (auction or dynamic-based pricing) with a substantial discount (up to 90%) in comparison with its on-demand price. However, spot instances can be terminated at any time with only two minutes of advance warning time. In 2015, AWS started to offer two modified pricing models for its spot instance: Spot Fleet and Spot Block. Following AWS lead, Google App Engine began to offer a cloud service platform (Platform as a Service or PaaS) for its customers to host their web applications within the current Google data centers in 2008. Its price model is very similar to AWS, but Google Cloud Platform (GCP) charges in per-minute base for Pay as you Go (PAYG). The underlying hypervisor of GCP for its PaaS is Kernel-based Virtual Machine (KVM) that was initially released by Qumranet in 2006. Later, it was acquired by Red Hat in 2008, but Red Hat was taken over by IBM in later 2018. In 2015, GCP also offered a discount (up to 80%) price or preemptible model for its cloud service to match AWS' spot model. In comparison with AWS and GCP, Microsoft Azure started its cloud business in Jan 2010. Its price models are very similar to both AWS and GCP. It has quickly captured a large market share according to Gartner's Magic [18]. Azure's virtualizing technology is built upon its own Hyper –V. Microsoft launched Hyper-V in 2008. In 2017, Azure

also followed the footstep of AWS and GCP to offer a "low priority" price model for up to 72% discount rate in comparison with "on-demand."

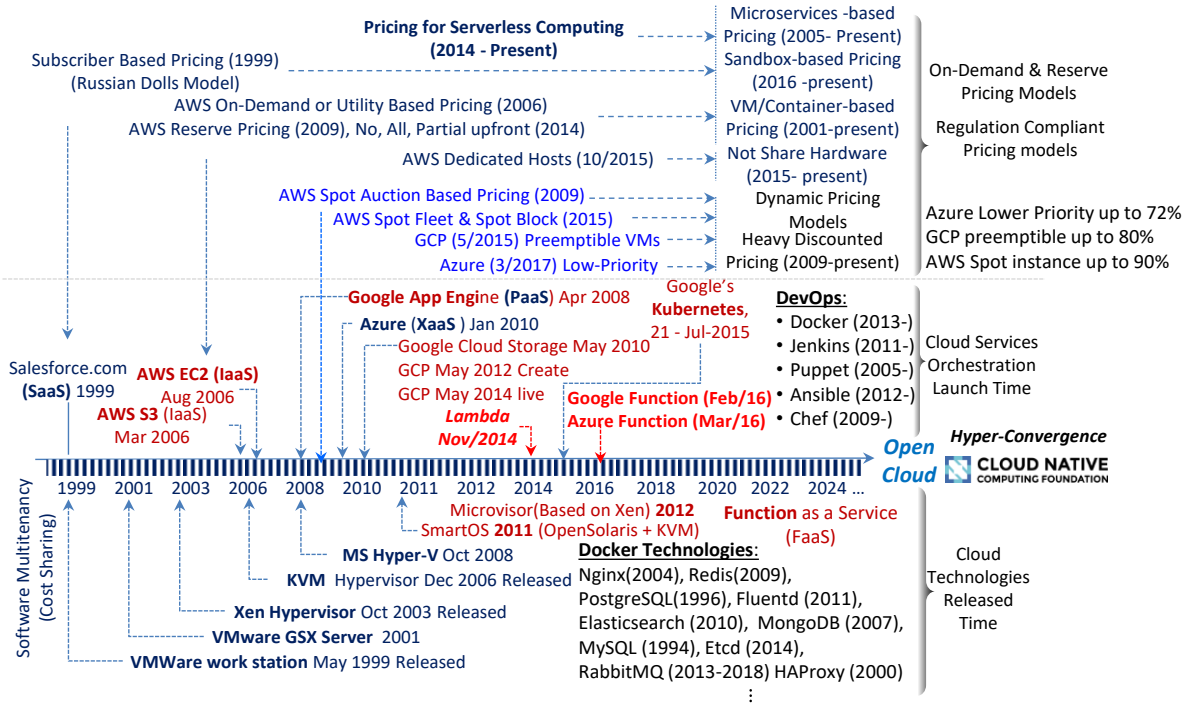


Figure 2—2 A History of Cloud Service Pricing Model and Enabling Technologies

Although the top three leading CSPs use three different hypervisors, many public CSPs adopt Citrix Xen, such as IBM Softlayer, Rackspace, GoGrid, Oracle VM for x86, Aliyun (Both Xen and KVM) and Virtustream's μVM (or Microvisor). Linode moved its VMs from Xen to KVM in Jun 2015 because it believes that KVM is 28% faster than Xen. However, the most popular hypervisor for the private cloud is still dominated by VMware, which is the first commercial hypervisor that was launched in 1999.

Some public CSPs, such as CenturyLink and Interoute, also adopt VMware to support their cloud business because VMware provides a comprehensive toolset that allows customers to manage their private cloud service efficiently. However, some analysts [105][106] argued if host applications are migrated to a public cloud, it will become too heavy and cumbersome. They also believe that one of the significant disadvantages of VMware is overpriced. One of the interesting observations is that most public CSPs adopt Xen hypervisors and the minimum billing unit of on-demand is per hour base. On the other hand, if CSPs adopt the KVM hypervisor, the billing unit

is reduced to a per-minute base. Some CSPs that adopt VMware often require cloud customers to have a long-term commitment to a cloud service contract. In general, virtualization technologies allow CSP to cut out the idle time of cloud data centers and improve cloud resource efficiency by 4-5 times. It enables CSP to reduce the significant amount of cloud infrastructure footprints. As a result, CSP can offer competitive cloud prices to its customers. Table 2—2 highlights the various price models and underlying hypervisors.

Table 2—2 Leading CSPs' Pricing Models and Supported Hypervisors

Name of CSP	Type Pricing Models	Initial offering Year	Minimum billing Unit/Cycle	Type of Cloud Service	Supported Technologies	Hypervisor Launched Year
Salesforce.com	Per-User Based	1999	Monthly	SaaS	Multitenancy	-
AWS	On-Demand	2006	Hourly	IaaS	Xen	2003
AWS	Spot Instance	2009	Hourly	IaaS	Xen	2003
AWS	Dedicated Hosts	2015	Hourly	IaaS	Xen	2003
Google App Engine	On-Demand	2008	Minute	PaaS	KVM	2006
Google Cloud Platform	On-Demand	2010-2014	Minute	IaaS	KVM	2010
Google Cloud Platform	Preemptible VMs	2015	Minute	IaaS	KVM	2010
Azure	On-Demand	2010	Hourly	XaaS	Hyper-V	2008
Azure	Low Priority VMs	2017	Hourly	IaaS	Hyper-V	2008
Softlayer (IBM)	On-Demand	2006	Hourly	IaaS	Xen	2003
Softlayer (IBM)	Bare Metal Cloud	2010	Hourly	IaaS	VMware	1999
Softlayer (IBM)	VMware Virtual Data Center	2016	Monthly/Yearly	IaaS	VMware	1999
Rackspace	On-Demand	2008	Hourly	IaaS	Xen	2003
GoGrid	On-Demand	2006	Hourly	IaaS	Xen	2003
Aliyun	On-Demand	2012	Hourly	IaaS/PaaS	Xen and KVM	2003, 2006
Virtustream (Dell)	Per User Based	2012	Monthly	IaaS	Microvisor (Xen)	2012
Joynet	On-Demand	2013	Minute	IaaS/PaaS	SmartOS	2011
Linode	On-Demand	2008	Monthly	IaaS	From Xen to KVM	2003, 2006
CenturyLink	Reserved Based	2011	Monthly/Yearly	PaaS	VMware	1999
Interoute	Reserved Based	2012	Monthly/Yearly	IaaS	VMware	1999
Oracle VM for X86	Reserved Based	2012	Yearly	IaaS/DaaS	Xen	2003

From a CSP perspective, we argue that discount pricing models would not support CSP's business profitability and sustainability economically. Instead, on-demand and reserved models

are the profit-driving forces for CSPs. The reasons to offer a discount price 1) CSP can fully utilize its spare cloud capacity. 2.) CSP can manage its cloud resources effectively for its cloud infrastructure lifecycle. 3) It can capture more customers' surplus values at a lower end of the pricing spectrum. 3.) It can become one of the marketing campaign tools for CSP to prompt other cloud services. 4) It can reduce customer churning by combining discount pricing with on-demand. Recently, AWS offered a modified version of spot instance: spot block and spot fleet, which is to combine on-demand and spot pricing. In comparison with pure on-demand, both models can save typically 30%-45% cost plus further 5% off for a non-peak time in a region. This is an excellent example to illustrate the AWS pricing strategy to reduce customer churning.

From a cloud customer's perspective, the reserved pricing model is to assure cloud resource certainty, and the on-demand pricing model is to accommodate customer's workload fluctuation with advantages of minimum provisioning time and speed to market. Currently, there are at least seven types of mainstream pricing models in the cloud market, namely On-demand, Reserved, Subscription, Discount (including auction), Code on Demand (CoD), bare metal, and Dedicated Host illustrated in Figure 2—3. These pricing models are mainly driven by cloud customers' utility values and market segments (Refer to Chapter 4 for more details). These models only show a practical aspect of cloud pricing in history. What is the theoretical aspect of cloud pricing in research?

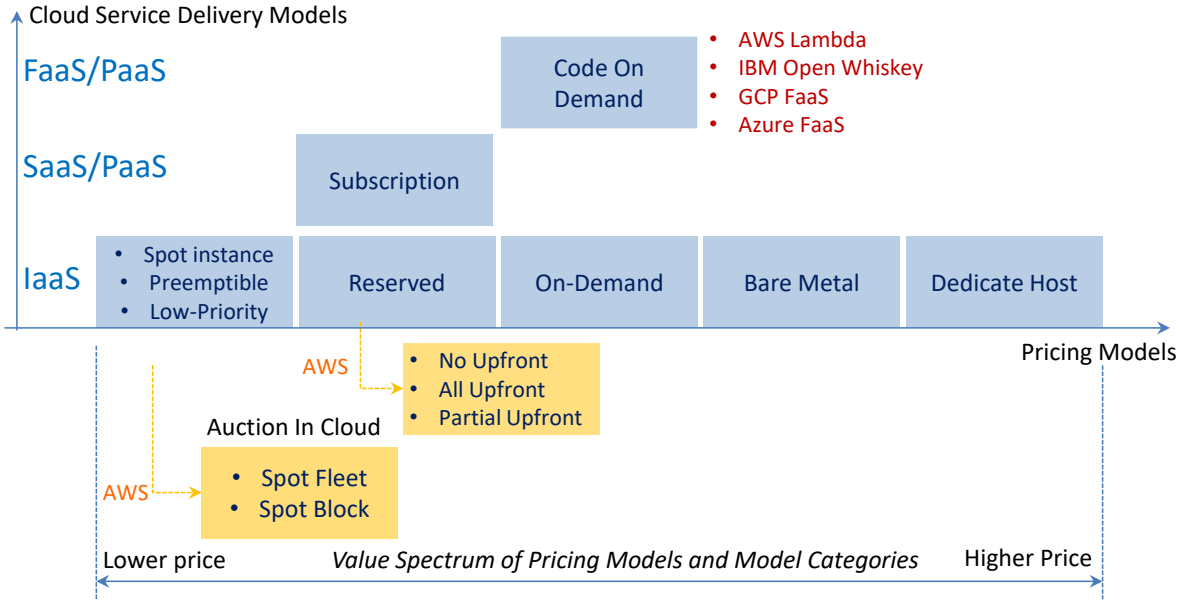


Figure 2—3 Summary of Cloud Pricing Spectrum in the Current Cloud Industry

2.2.2. Multiple Roots of Cloud Pricing Models in Research

The aim of examining various pricing theories is to clarify multiple roots of cloud pricing theories. By tracing down the historical roots of various cloud pricing models that were proposed by many researchers (more details in the following Section 2.4), we can see that the origin of cloud pricing models does not come from single but multiple threads. The current term of the cloud pricing model is an amalgam of different sources. On the basis of reviewing more than hundreds of research papers between 2008 up to present, we can identify possible four primary roots of cloud pricing (as shown in Figure 2—4): Utility-computing, Network computing, CSP's profit-driven, and cloud customer performance orientation. This historical tracking suggests two possible criteria to classify various cloud pricing models. One is to classify pricing models by its historical roots, and the other is to carve (cloud pricing) nature at its (economic) joint [107]. This study will exhibit the taxonomy of the cloud pricing models based on economics and value theory because it will be not only in align with economic theory, but also help many decision-makers to comprehend a value proposition of each pricing model.

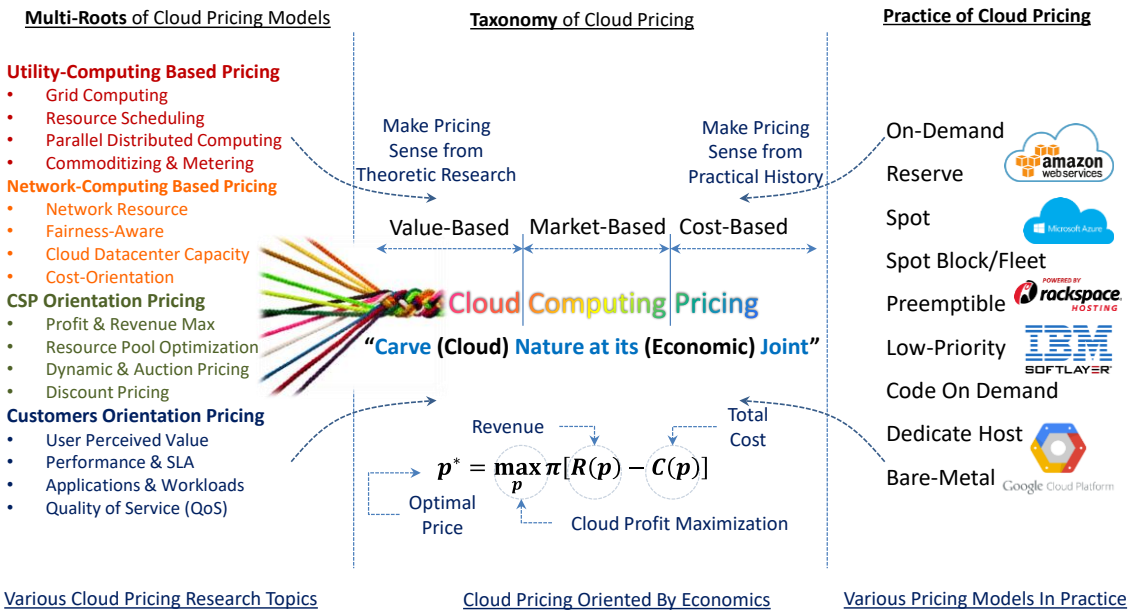


Figure 2—4 Multiple Roots of Cloud Pricing Models

2.2.3. Key Terms, Strategies, and Relationship of Pricing Models

From practice to theory, we have introduced many terms regarding the cloud pricing model. However, the meanings of some key terms and their relationship are still vague in terms of cloud

pricing contexts, such as price, pricing, pricing scheme, pricing model, pricing structure, pricing category, pricing strategies, value, and customer benefits. These terms and their relationships are essential for the following taxonomy and survey.

The term price is an estimated value or a value tag of cloud service (e.g., $1.00/per hour). Pricing is to give an estimated value based on a value proposition. The pricing scheme is a price plan or cloud service package with a pricing tag (e.g., AWS c4.larg instance consists of 3.75GB-RAM, 8-ECU or EC2 Compute Unit, 2-vCPU or virtual Central Processing Unit, Linux-OS, and is marked as $0.10/per hour at US East Ohio data center in April 2019). It may be considered as a price configuration. Some CSPs allow cloud customers to create their own pricing scheme by setting a range of standard prices. Pricing model (e.g., on-demand or reserved) is a simplified description that is often defined by a mathematic function for CSP's profit maximization (e.g., $p^* = \max_p \pi[R(p) - C(p)]$, where π is a profit, p is a price, $R(p)$ is total revenue and $C(p)$ is a total cost). Pricing category is a group of pricing models that has some common characteristics, while pricing strategy is a blueprint by coordinating various activities to achieve a long term business goal (e.g., A strategic goal is to achieve a 20% revenue growth in next five years). If the pricing scheme is an abstraction of various prices of cloud components, then the pricing model is an abstraction of pricing scheme, and pricing strategy can be considered as an abstraction of pricing model. They are all dependent on a set of value propositions for a purpose to deliver cloud customers' benefits (Refer to Figure 2-5).

The term "value" means how much worth to an agent for an object. It is measured by a unit of utility [120] (worth, satisfaction, happiness and subjective experience). Fundamentally, value concerns things are good or bad in a successful and efficient sense [22] [23]. To this extent, it can be further articulated into three types of good values: 1) "Good to have" (e.g., a pricing strategy aims to consolidate good customers' experiences of cloud services), 2) "Good to do" (e.g., the strategy drives the customers' value proposition of willingness to pay, which focus on new values), and 3) "Good to be" (e.g., a strategy is to simulate customers demand to migrate more workloads to off-premises). By delivering various "good" values of cloud services, customers are willing to pay (W2P) for their service benefits and CSP will get a profit reward from its cloud business. It means "value co-creation" [124]. We can briefly illustrate the relationship with all these key terms in Figure 2—5. The aim of having three types of "good" is to know how to handle all the factors

of pricing model so that a cloud decision-maker can know which factor is more important than others.

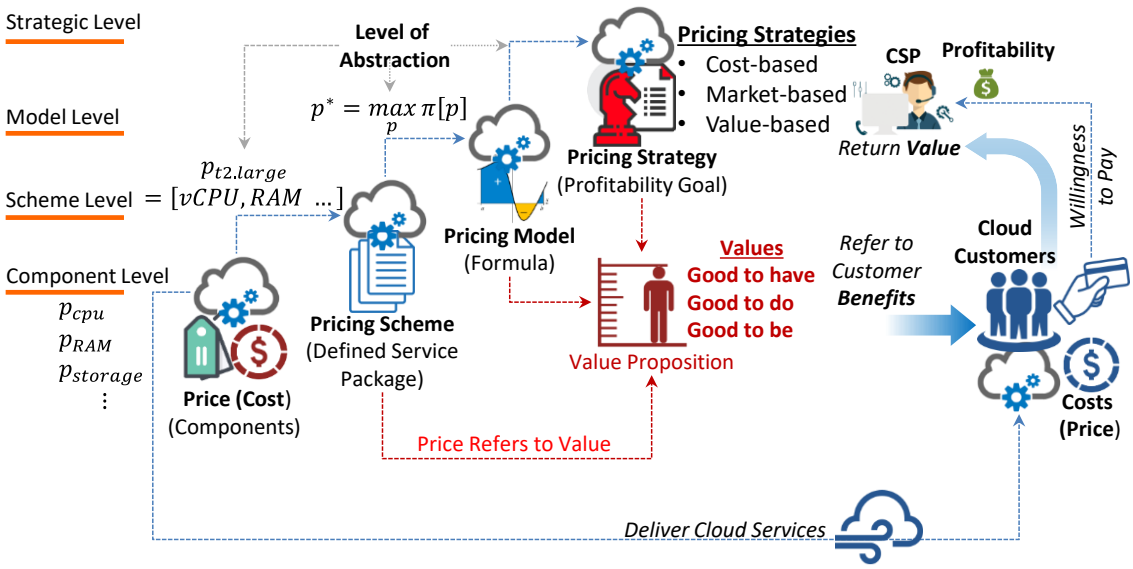


Figure 2—5 Relationship of Key Terms for Pricing Models and Value Proposition

2.2.3.1. Value-Based Pricing

In comparison with other pricing strategies, value-based pricing is much subjective. It might not be necessary to reflect on a market price and service costs. A typical example is perception-value, which is based on the customers' perceptions of what is expected in comparison with what is to be delivered by a CSP. The common term of perceptive value is value for money that is the ratio between the worth of a cloud service and a price to be paid [24]. According to Sheth et al. [25], customers perceived values have five dimensions, namely functional, conditional, social, emotional and epistemic values. The final decision of customer choice is a function of multiple perceived values. The main benefit of value-based pricing is that it provides competitive advantages to capture a wide range of cloud services' values [26], such as emotional and epistemic. However, it is quite challenging to be constructed because "perceived values" are the primarily measured satisfaction of the individual customer. With the B2B type of service [27], it is even hard to detect end-users' satisfaction directly. Instead, the perceived values could be influenced by an indirect person, such as a manager's or decision maker's perception

In principle, the value-based pricing emphasizes the measurements of customer's experience, satisfaction, and expectation. It includes both intrinsic values [9], e.g., CPU, RAM, bandwidth and extrinsic (or instrumental) values [10], which are determined by the relationship about something, e.g., Pay as You Go (PAYG), 24X7 supports, burstable CPU, resource auto-scaling, etc. (See Chapter 3 for more details). The value-based pricing is often applied to innovative cloud service features and some new niche market segments. By a similar line of reasoning, we can extend the value-based criteria to both market-based and cost-based pricing. Consequently, we can form a 3 × 3 matrix as the classification criteria to differentiate various cloud pricing models listed in Table 2—3

Table 2—3 Classification Criteria Matrix for Cloud Pricing Models

Pricing Strategies	Value-Based Pricing	Market-Based Pricing	Cost-Based Pricing	Target Values
Good to Have (GH)	GH for Value-Based	GH for Market-Based	GH for Cost-Based	Consolidate Current Values
Good to Do(GD)	GD for Value-Based	GD for Market-Based	GD for Cost-Based	Grow New Values
Good to Be (GB)	GB for Value-Based	GB for Market-Based	GB for Cost-Based	Identify Future Values

2.2.3.2. Market-Based Pricing

"Market-based pricing" is driven by the equilibrium of all customers and CSPs [28]. The market environment will stabilize the market price, which is price equilibrium due to supply and demand. We can use "Freemium" as one of the examples to illustrate the idea of market-based pricing, which it becomes popular due to rising FaaS (Further details in Section 2.4.4) "Freemium" is to give away a product with basic functionality or features for free to gain the market share [29]

The primary purpose of Freemium is aiming to convert free customers into premium buyers by giving away just enough values for initial taste so that it can attract regular customers. That is why the word "Freemium" is the combination of two words of "Free and Premium." "Freemium" is one of the pricing models for many CSPs, such as AWS, GoGrid, SoftLayer, Dimension Data, Microsoft Azure, ElasticHosts, and Dropbox, to implement their market-based pricing strategy. The market-based pricing takes consideration of two kinds of impacts on the pricing. One is price-sensitive and the other is the competitiveness of a market for similar services. Practically, CSP

[9] Intrinsic Value – A value can be isolated by its own
[10] Extrinsic Value – A value is dependent on others, or instrumental value

may adopt different pricing models to implement its business strategy, such as classic feature-limited freemium (such as AWS and Dropbox), Free trial period (such as Azure), unlock the capped speed or bandwidth or unique features (such as mobile apps and gaming), free software and premium service support. Moreover, these models can be measured by various metrics. Marius F. Niculescu et al. [30] highlighted four different measurements, which are features, quantity, quality, and period. These models can attract many high-end customers and get much valuable feedbacks from a large number of audiences for a CSP to improve its services. The criteria to classify this marketed-based pricing can be summarized as to be competitive in response to a market price due to supply and demand.

2.2.3.3. Cost-Based Pricing

Although market-based pricing is common for many retailer businesses, most of the enterprises and government agents with on-premises cloud infrastructure often adopt cost-based pricing because it is much easier to be comprehended from a decision-making perspective. One of the primary reasons to use this pricing strategy is it is concrete and tangible. There is no other interpretation. It is also known as fact-based pricing. Despite the fact that many pricing experts emphasize value-based pricing [31] [32] [33], the cost-based pricing is still common because it can help decision-makers to set a baseline price to charge customers so that they can at least cover Capex. Moreover, cost-based pricing can articulate a unit cost and provides a reference point for benchmark comparison. It becomes one of the managerial tools for many decision-makers to drive CSP's business performance. Lastly, the components of cost are the essential element of Cost and Benefit Analysis (CBA) so that a decision can be made much realistic [34] and a cloud price can be validated internally.

In practice [35], value-based pricing is often far less than the other two pricing strategies, and market-based pricing is the most popular strategy and followed by cost-based pricing, as shown in Figure 2—6. This result indicates value-based pricing is much more challenging to be applied due to a value estimation of cloud customers' experiences, satisfaction, and perception. T. Nagle et al. [10] proposed a practical solution of value metrics, which consists of six actives or value cascade to implement value-based pricing. They are value creation, value communication, price structure, pricing policy, price setting, and price competition. With clarification of three pricing strategies and some critical terms of cloud pricing in upfront, the taxonomy of cloud pricing can be developed.

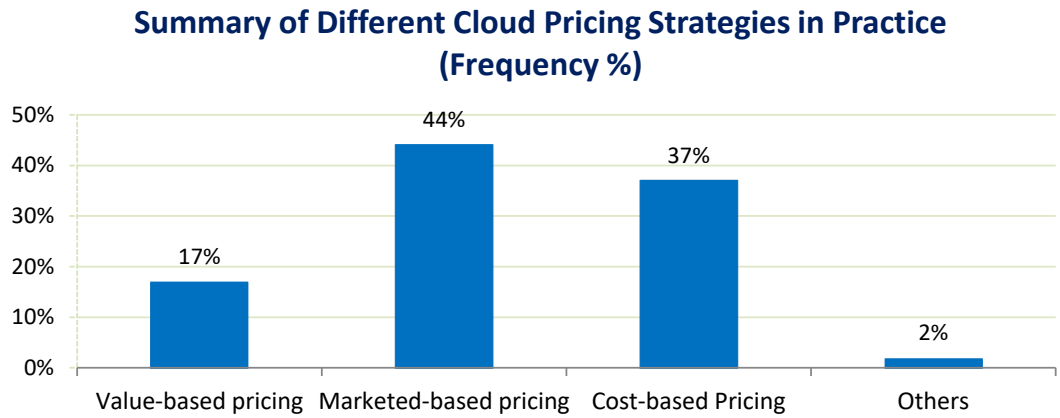


Figure 2—6 Adoption of Pricing Strategies in Practice Across All Industries [36]

2.3. Taxonomy of Pricing Models

Based on the root of three pricing strategies, we can further map onto 60 different pricing models that are determined by four factors of value, fact, supply, and demand, which we can build a comprehensive framework of taxonomy that includes 9 different categories that are formed a 3×3 matrix as shown in both Figure 2—7 and Figure 2—8 Each category of pricing consists of between 3 and 6 pricing models except retail-based pricing models. From Section 2.3.1 to Section 2.3.9, this chapter will first define each category and then will explain why some models have been adopted by CSPs and others not. Finally, this chapter will link each category to today's cloud pricing practice

Notice that it is possible to carve various pricing models at different joints. It may lead to one price model to be mapped onto different categories and different strategies. It is dependent on many factors, such as a business strategy, investment budget, the expertise of cloud technology, a competitive market environment, and targeted customers. Ultimately, it is dependent on a value proposition. In practice, we can combine various pricing models to form a new pricing category and to achieve a particular tactical object. This taxonomy, together with three pricing strategies and a 3×3 value matrix defines the conceptual framework for cloud decision-makers to comprehend cloud price models systematically.

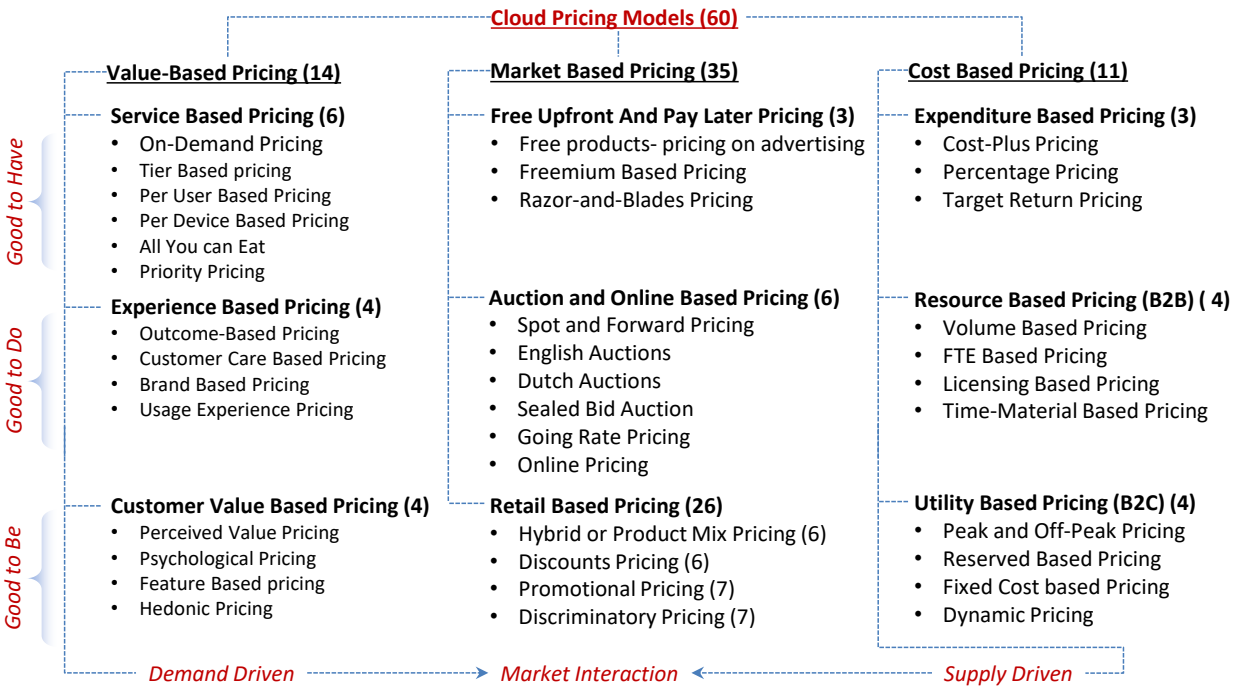


Figure 2—7 Taxonomy of Overall Cloud Pricing Models

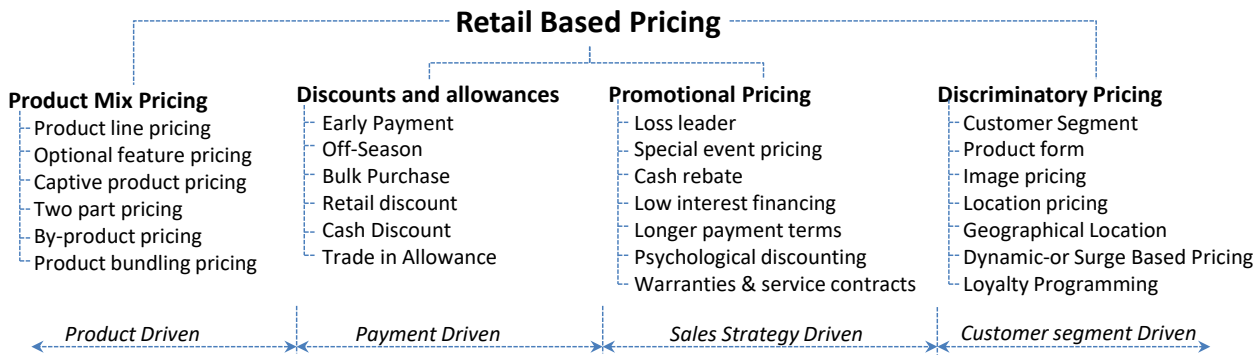


Figure 2—8 Taxonomy of Retail-Based Pricing

2.3.1 Service-Based Pricing

Service-based pricing category is to emphasize the value of "Good to have" for cloud customers. In comparison with the other two value-based pricing categories, the value of "good to have" focuses on value consolidation for the cloud services. Many CSPs of SaaS adopt service-based pricing models, such as Salesforce.com and Azure. It can also be considered as incentive-based pricing because this pricing category could be determined by a client's business revenue, cost-saving, and early project delivery. The advantage of this category of pricing models is their values

can be identified and predicted. The value of "good to have" can be incremental. There are six different models of value pricing: on-demand, tier-based, per-user based, per device-based, all you can eat, and priority-based pricing. The value measurement of these models may be dependent on a Service Level Agreement (SLA) [63] [91]. Although service-based pricing is closely associated with performance pricing due to SLA measurement, the former focuses on the pricing of service contents while the latter aims at the pricing of the performance required.

The concept of service-based pricing could also be mixed with resource-based pricing because both categories of pricing may involve some components of intangible inputs and outputs. However, the service-based pricing focuses on value-added service, while resource-based pricing emphasizes the requirement of various inputs. The typical example of service-based pricing for cloud service is "on-demand" or PAYG, which is one of five essential characteristics of cloud service [197].

2.3.2. Performance-Based Pricing

This pricing category can be distinguished as the value is "good to do." It is measured by customers' performance experiences, such as the specified reliability of a cloud service or utilization rate of a limited resource (e.g., cloud infrastructure or data center capacity). The aim of these models is to sell the new service values to customers for performance requirements, such as end-users response time, network throughput, latency, security, and scalability. To some extent, it may also be considered experience-based pricing. According to M McNair's definition [40], "Performance-based pricing is an arrangement in which the seller is paid based on the actual performance of its product or service."

A typical example of performance-based pricing is online advertising payment, which is dependent on the measurement data, such as the number of clicks or purchases [41]. Other applications include telecom services (such as multi-party video conferences, mobile apps, satellite connectivity, etc.), in which the service prices rely on its specified performance metrics. This pricing category is often connected to the customer's business outcome. The basic idea is to make sure that a CSP's services meet the customer's business objectives or value. The reward of this model is that both parties' values are aligned. By doing so, the CSP will not undercharge the pricing, and a cloud customer will be given the performance guarantees for the services. The advantage of these models can become "win-win" pricing models and be fair to both parties. From a customer perspective, this model shifts the uncertainty risks to a CSP. However, not every

performance metric can be quantified or determined. Sometimes, the performance metric is quite complicated. For example, how to determine the length of the period for the number of clicks for one online adverting camping? Often, the advertising campaign time may take longer than what was initially expected. In practice, the performance-based pricing models can be subdivided into four different models based on customer's experiences of "good to do." They are outcome-based, customer care-based, brand-based, and usage of experience-based pricing. In comparison with other categories of value-based pricing strategy, the performance-based pricing is practical because of its definable performance metrics. In a cloud practice, many B2B cloud services emphasize on performance-based pricing, which a CSP offers a guarantee performance, such as five-nine service reliability or 20 Gigabit/s network throughput and in return to charging a premium price.

2.3.3. Customer Value-Based Pricing

This category of value-based pricing consists of four pricing models, namely perceived-value, psychological, feature, and hedonic based pricing. A customer's core value is the main reason to build various price models. If the customers believe the cloud service value offered by a CSP is "good to be," they will be willing to pay (W2P) for it. These four models are constructed by the context of perception, psychology, sociology (broad environment) and economics (utility). The primary advantage is that it allows a CSP to maximize its business profit and lead the cloud market. The main challenge is how to define the value metrics by measuring customers' subjectiveness value for "good to be." In comparison with other models, both feature and hedonic pricing, Chapter 3 can be established if the historical dataset is available. These models can be effectively applied to an ever-changing environment in terms of new cloud features (characteristics). However, not every feature of service would be "good to be" for every customer. As a result, a decision to select cloud service features corresponding to the charging price could become a challenge from a CSP perspective. Kilcioglu and Rao [117] observed one of the possible solutions was to modularize cloud resources and build a relationship between cloud services and customer value metrics for future growth. This solution has been implemented by many CSPs.

2.3.4. Free Upfront and Pay Later Pricing

Due to intensive market competition, many CSPs adopt "Free upfront and pay later" pricing model. The idea is to leverage free products with minimum features so that the pricing model can capture more customers and make the profits from premium customers. There are often three

types of models, namely free products-pricing on advertising, Freemium, and Razor-and-Blades pricing. With free product pricing on the ads model, it can stimulate customer's demand, and customers can enjoy free products. The bad news for customers is that they could waste a lot of unnecessary time to try various free products. Moreover, this model requires a sizable market from a supplier's perspective. If the market size is not large enough to offset the cost of free products, the pricing model is unsustainable. For the freemium model, there are four types of sub-category models: 1) Classic feature-limited freemium (AWS and Dropbox adopt this model). 2) Free trial period (MS Azure and Oracle cloud services). 3) Free software and premium service support (Red Hat Linux), and 4) Unlock the capped speed or bandwidth or unique service feature (mobile apps, gaming, and pay-TV services). These models are pricing four different values, which are quantity, period, quality and service features. The critical issue is how to draw a line between free and premium services. Recently, AWS began to offer Lambda service or FaaS, which is one of the freemium services in term of quantity (execution times or a number of clicks and memory size/per month)

Razor-and-Blades model is similar to freemium, but the main difference is Razor-and-Blades emphasize the concept of regular and consumable components. For example, a provider may give away or charge a minimum price for the initial or not-consumable element, such as a printer but charge a high premium for a regular and consumable replacement component, such as printer cartridges. The main advantage of this model is it can optimize the product prices and increase sales and maximize the business profits by redefining different values of product components. However, not every product can be divided into "Razor" and "Blades" Moreover, with the intensive market competition, the provider may risk recovering the "Razor" cost due to losing the returning customers. From a value perspective, these market-based pricing models are "good to have" to consolidate a CSP's market share. Now, many leading CSPs start to offer this pricing model for their Function as a Service (FaaS), such as IBM Openwhisky, GCP, and Azure function services. FaaS based pricing is one type of Free Upfront and Pay later pricing.

2.3.5. Auction and Online-Based Pricing

2.3.5.1. Auction Pricing and Auction in Cloud

Auction-based pricing is that the auction mechanism will decide the pricing. Asunción Mochón [46] stated: "Auction is a market mechanism, operating under specific rules, that determines to whom one or more items will be awarded and at what price." The reason for the auction-based

pricing is that the market price of some products, such as artworks, antiques and certain rights (radio spectrum licenses), would be best to be settled via pricing bidding mechanisms. Today, numerous products and services are under a hammer from inexpensive items sold on the internet (eBay) to a billion-dollar mobile spectrum license. Many commodity products, property, and financial bonds are included. AWS also places its EC2 and S3 under its auction bidding rules.

There are some pros in term auction-based pricing: The speed of the auction is relatively fast. There are no backward and forward processing steps. The price is also very transparent, which the bidder only pays the increment cost at each bid. Moreover, it is fair to all bidders or players who obey the auction rules. The auction process is straightforward and direct. The limitations of the auction are: For a bidder (or customer), they have very little time to think during the bidding process. Subsequently, it is the price that may be overbidding on the real value of goods. Under the auction theory, there are different types of auctions based on the design criteria. Lawrence M Ausubel [47] listed about 13 different kinds of auctions:1) Clock auction, 2) Combinatorial auction or package bidding, 3) Dutch auction (Open Descending), 4) English Auction (Open Ascending), 5) First Price Auction, 6) Second Price Auction, 7) Pay-as-bid action, 8) Revenue Maximization or optimal action, 9) Simultaneous ascending auction,10) Uniform-price auction,11) Vickrey auction (Second Price Seal-Bid Auction),12) Vickrey-Clarke-Grove (VCG) mechanism, and 13) Winner's curse. These are different auction forms.

The popular auction pricing models can be categorized into four models: Spot and forward pricing, English Auction, Dutch Auctions, and Sealed-bid Auction. This chapter only focuses on a few auction models that are closely associated with the cloud market. For example, AWS has adopted a modified spot and forward pricing since 2009. The term "spot" literately means the value of an asset at the right moment of settling based on English auction. It is derived from a commodity market. "Modified" means that AWS spot instance is not a real spot price because AWS reserves its right to toss or terminate your bided instances at any time by providing two minutes warning time in advance. Currently, the only CSP or AWS provided the spot instance for public cloud customers. In 2015, AWS offered two modified version of spot instances, namely Spot block, and Spot feet to exploit more customer's surplus-value. With spot instance, a customer only bids for one instance. Spot block means the customer can bid for an instance that can lock in a finite number of continuous runtime hours (from 1 to 6 hours). For the Spot fleet, AWS allows a customer to bid multiple spot instances from a spot instance pool. AWS also allows customers to mix with different pricing models (e.g., on-demand and spot instance) to form a specified

computing capacity, such as 10 VMs that consists of 8 on-demand instances and 2 spot instances. The auction-based pricing model can be considered as designing for a niche and growing market, such as big data analytics workloads. Economically, the aim of the spot pricing model is similar to other discount pricing models, such as GCP's preemptible or Azure's low-priority VM, which is to capture more customers' surplus values at a lower end of the cloud pricing spectrum.

2.3.5.2. Online Pricing

In contrast to offline pricing, the meaning of "online" is the purchasing goods can only be processed via the Internet and cannot be handled offline or in a physical store. However, some online retailers may also offer both online and offline purchasing prices for customers, but the offline price could be higher than the online one. For example, Officeworks provides both online and offline prices, but the offline price is sometimes higher than online.

The upside of online pricing is it can instantly reach a vast number of customers for a provider. The purchase transaction can be made very quickly via an electronic transaction. There are no extra handling expenditures except postage costs. It is much convenient for a customer to do online shopping and make an easy for the customer to compare different online pricing with different online suppliers. Overall, online pricing enables customers to do the shopping and achieve at least six benefits: "shopping at a finger-click," saving time, competitive pricing, a wide range of goods, no time pressure for shopping and reading product information details, and various brands and models to be selected. The downside of online pricing is high risks of security and privacy issues, lack of or no significant discount, fraud in online pricing and the extra cost of goods delivered. From a CSP perspective, it can leverage online information via a recommendation system to tail cloud services for a personalized price or price discrimination. As a result, the CSP can improve its both revenue and profit margin.

2.3.6. Retail-Based Pricing

By its name, the retail-based pricing models are based on a small quantity that consumers buy from physical locations or retail outlets (such as, discount shop, warehouse, factory outlet, shopping malls, petrol station, department stores, supermarket, Sunday market, etc.) By and large, the retail providers sell products in a small quantity. It is mainly business to customer (B2C) type of pricing model. However, some models are also applied in B2B. There are at least four subcategories of pricing models: product mixing, discounts, and allowances, promotional, and discriminatory pricing. Altogether, retail-based pricing has a total number of 26 models. Each

pricing subcategory has a different orientation, as shown in Figure 2—8, which the products nature drives the product mix pricing, the payment option drives the discounts and allowances pricing, the sale strategy drives the promotional pricing, and the customer segment drives the discriminatory pricing.

2.3.6.1. Product Mix Pricing

This pricing model category is to mix or combine with different types of pricing models in different ways. Providers can depend on customers' usage patterns to combine different pricing models. The standard practice for cloud services is to combine both on-demand and spot instance pricing models to accommodate both predictable and unpredictable workloads [39]. There are six types of product mixing models, namely product line, optional feature, captive product, two-part tariff, by-product, and product bundling. The primary focus on this subcategory of pricing is the relationship of different products, which is how to mix various services to achieve the maximum profit by consideration of limited resource capacity, perishable assets, marginal cost and an optimal mixture of multiple products.

The benefits of these models can boost sales, generate extra revenue or profits and meet various demands or market segments. However, the main disadvantages of these models are some customers may feel the frustration of trapping into a cost black hole. Others may decide not to buy at all. It may create a backlash among some premium customers and lead to a bad reputation for service providers. It may also increase the provider's operational costs. The bottom line is how to make a rational decision on pricing that can reflect customers' demands by different segments. Recently, AWS had implemented this type of pricing model in 2015, which is called "Spot-Fleet." The distinct advantage is that it can reduce the customer's churning rate and increase sales revenue and profits.

2.3.6.2. Discounts and Allowances Pricing

Price discounts and allowances are two techniques for a firm to response fluctuation conditions due to market dynamics. The term discount represents a firm to give a pricing reduction because of product promotion, off-season, cash payment, bulk purchase, display, and bundle, wholesale, and two-part tariff. This technique is applied to many perishable services. Cloud Computer resource is one of the perishable assets. AWS had a few price reductions between 2006 and 2014 [42]. Allowance pricing is another type of price discount, but it is mainly designed for wholesale customers or commercial clients or SME. Overall, this subcategory of pricing models has six

kinds of common discount and allowances pricing models, which are early payment, off-season, bulk purchase, retail discount, cash discount, and trade-in allowance.

The goal of this subcategory is "payment-driven" to improve net present value (NPV), which is to increase the return of net cash flow. The benefits of these pricing models are to reduce the stock inventory or to improve the capacity utilization rate, especially for perishable assets, like cloud resources. The main disadvantage of these models may reduce the profit margin and do not have a brand identity. Currently, all three leading CSPs are offering a price discount, such as spot, preemptible, and low priority for the number of reasons presented in the above Section2.2.1.

2.3.6.3. Promotional Pricing

Promotional pricing is one of the sales strategies, which is to give a discount within a specified period. "Most product management teams will create and agree upon a seasonal promotions calendar for their business. The calendar plans out the flow of promotions over a year and is used as a framework that ensures that the available product is sufficient to meet customer demand and maximize business opportunities. Promotions help generate demand and provide for immediate cash flow into a business. Likewise, promotions can help stimulate demand for slow-selling products and so can help reduce product over-stock" [34]

The obvious reward is to increase sales and minimize stock levels [44]. The drawback is that it will drag down the overall profit margin. There are seven different pricing models to boost sales, which are a loss leader, special event, cash rebate, low-interest financing, longer payment terms, warranties and service contracts, and psychological discounting. The primary focus of this pricing subcategory is the sales-driven. One of the typical examples is a laptop sale with a cash rebate for a particular model of the laptop. Recently, GCP has started to offer a promotion price for its cloud Tensorflow Process Units (Cloud TPUv2) for US$4.50/per hour [112] in comparison with TPUv3 with an $8.00/per hour. The price is substantially low in comparison with a regular price.

2.3.6.4. Discriminatory Pricing

Discriminatory pricing means that the pricing model is charging different prices to different customers for the same services. If we look from a value perspective, it is a customer value-based pricing strategy to charge each customer at the maximum price according to the customer's perceived value, which is the price that a customer is willing to pay. Based on the classification of the microeconomic theory [8], if it is the 1st degree of pricing discriminatory, the price is usually dependent on one-to-one negotiation, such as property sale (in private sale). It often

requires a lot of effort to capture the customer's maximum value. It is less likely applied to a commodity product.

If the price discriminatory (or price discount) is dependent on sales volume, this is called the 2nd price discrimination. The typical example is a bulk-purchase discount in comparison with a single purchase. It is a common practice for wholesale. If the price charge is based on the specific group of people in society, such as senior citizens, students, it is the 3rd degree of price discriminatory. For instance, Microsoft charges a student license for MS office package. If we combine different types of price discriminatory, we should have various price models in practice.

Overall, they are seven different types of pricing models: customer segment, product form, image, location, geographical location, dynamic or surge-based, and loyalty programming pricing. The main idea behind this subcategory is customer segmentation, which is to design different pricing models for various groups of customers. Amazon segments its customers by mixing operational revenue streams [45], as shown in Figure 2—9. This subcategory of pricing models does not only allow a CSP to boost its sales but also to maintain the profit margin. The flip side of these pricing models would increase sales costs, which will ultimately increase the investment risks. The criteria of model classification are two measurements: market segmentation and value principle of "Good to be" to create new values for CSPs. In the cloud industry, the practice of discriminatory pricing is pervasive, especially for cloud storage services. "Bulk -selling or purchase" that is 2nd order discriminatory is a typical example. AWS S3 has a bulk-selling price.

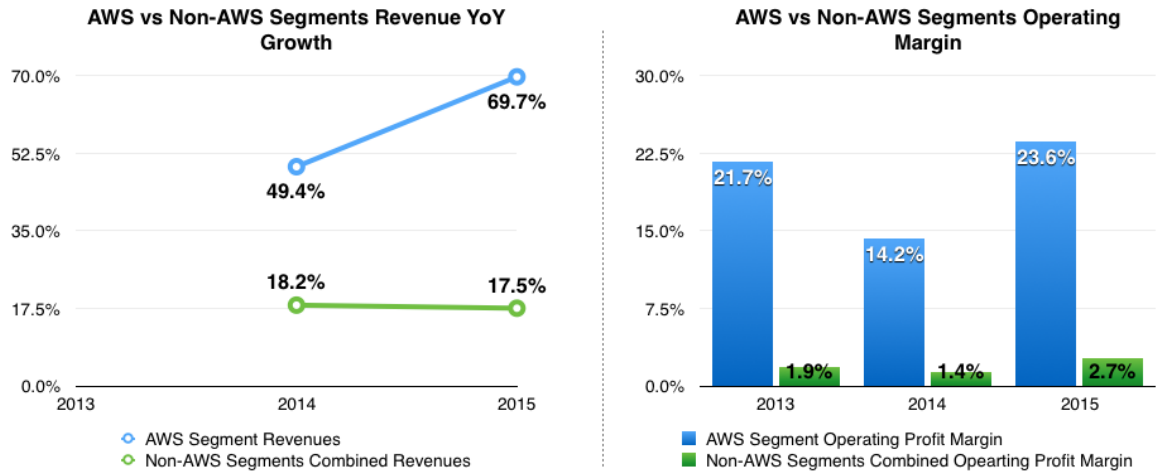


Figure 2—9 Amazon Segmentation of operational revenue [45]

2.3.7. Expenditure-Based Pricing

Expenditure-based pricing means every price model is derivate or built up from the center component – a unit of "cost." In this category, there are three types of pricing models, namely: cost-plus, percentage, and target return pricing. The primary driver behind this category is all price values proportional to the particular percentage of the total cost.

The benefits of these types of models are that a CSP knows a targeted-return. They are very concise, more straightforward and quick to be constructed. They can guarantee the profit bottom line, at least from a modeling perspective. However, these models ignore customer values and market supply and demand. Subsequently, these models may result to be either overestimating or underestimating the market price. Moreover, if the expenditure (cost) item is inaccurate, it would lead to the wrong pricing. Furthermore, the end to end (E2E) or the total expenditure for many large enterprises and government agents are not transparent. It is quite often that one cost item has been accounted multiple times. If so, it leads to overestimating a price for offering services. As a result, larger firms or enterprises could lose many business opportunities. However, if some cloud customers have some special requirements, such as regulatory compliance for their running business applications regarding cloud infrastructure, expenditure-based pricing models are good to have. In 2015, AWS released a new pricing model: a dedicated host to meet customers' compliance requirements. IBM has had a similar price model, called "bare metal," to eliminate the "noisy neighbor" effect. All these models are driven by cloud expenditure (or costs). This kind of pricing model may appear to be contradictory to the concept of cloud, but it is fit into particular business requirements – regulatory compliance, a high degree of security control, streaming applications, and dedicated computing power.

2.3.8. Resources-Based Pricing

Instead of pricing on cost account, resource-based pricing focuses on a consumption base. Sometimes, it might also be considered as activity-based pricing (costing) [121]. We classify resource-based pricing as one of the categories for the cost-based strategy because they have some common properties that are associated with the expenditure components. However, not all resource consumption costs money. Some natural resources are free. For example, the natural resource of solar or wind power does not cost any money. The resource-based pricing emphasizes on scalability. Many cloud services are built on resource-based pricing. Chen [122] found the cloud market or customers have a stronger preference for a particular CSP or a CSP can offer higher SLA than its competitors. The CSP is more likely to adopt the resource-based- pricing.

Resource-based pricing is common for the services industry. Traditionally, there are many service industries that adopted resources-based pricing models, such as e-commerce, airline, travel and leisure, recreation and entertainment, healthcare, and education. Resource-based pricing is also adopted by the IT industry, especially for IT outsourcing purposes. Resource-based pricing aims to offer a better method that allows customers to consume and deploy the scalable resources both efficiently and effectively.

This category of pricing emphasizes resource scarcity [102]. There are four types of resource-based pricing, namely, Transaction-based, FTE-based, Licensing-Based, Time-Material Based pricing. We can roughly differentiate this category of pricing models by criteria of "Good to do." Softlayer and VMware recently launched the "VMware virtual data center." It can be considered as one of the resource-based pricing because it includes all resources of cloud service, even including archive storage resources.

2.3.9. Utility-Based Pricing

Ruparelia, Nayan B [39] defined the term of a utility pricing model as: "Utility models are metered price models whereby your usage of the service is monitored, and you pay accordingly." His further explanation is that the origin of the model was "from the price plans that utility companies have adopted, they are characterized by regular payments, often monthly, to the cloud service provider."

The term of utility has serval different connotations. 1) From a computer software perspective, it means that the software can perform multiple specified functions. For example, utility software (iOS or Windows) can be utilized to perform the tasks of monitor, mouse, printer, and disk driver. 2) Another meaning utility is very close to the utility function, which is utilization rate for a certain amount of capacity. 3) From a public service perspective, it means an incumbent service provider can provide public services, such as telecom, electricity, gas, water, public transportation, which these services are essential to the modern society. 4) The economic term of utility is that the person receives satisfaction or pleasure for consumer goods or services. The original meaning of "utility" was coined by Bentham [48], which means the principle of utility or usefulness that is "greatest happiness for the greatest number of people."

For the category of utility-based pricing, the meaning of utility is similar to the metered price for public services. The benefit of utility-based pricing is that every individual can access the

cloud service directly via a credit card for infinite scaled resources without a prerequisite condition, upfront Capex. The flipside is that it is not a good idea to commoditize some new or innovative cloud service features by using this model. Nevertheless, this type of pricing model provides the value of "good to be" for cloud end-user because of OpenStack [123] development. According to various business requirements, usage time, resource commitment, customer segments, and payment types or different workload patterns, utility-based pricing can have different pricing models, namely, Peak and Off-Peak and fixed cost-based pricing. Chen et al. [122] argued that if cloud market demand is less volatile, cloud customers would prefer the resource-based pricing. In contrast, if their demand is highly volatile, they would prefer to utility-based pricing.

2.3.10. Summary of Pricing Models Classification

From both Figure 2—2 and Table 2—2, we can find that service-based pricing, especially on-demand, per use-based and tier-based pricing models, became common pricing models widely adopted by many CSPs. The aim of these pricing models is that they can reflect the cloud characteristics of both scalability and "on-demand" (or Pay as You Go). If we look back 40 years' computing history as shown in Figure 1—1, we can see that billing method is moving from "Pay As You Make" to "Pay As You Use" or "Pay As You Can" and the delivery model is moving from "Big Iron" to "FaaS" and pricing model is moving from hardware base to functional base. Altogether, a pricing strategy is moving from cost-based to value-based one. However, it does not mean cost-based pricing will disappear. They could co-exist with various new types of pricing models based on the computing technology adoption lifecycle [15].

As we have also shown in Figure 2—3, there are approximately seven cloud pricing models or model categories offered by leading CSPs at the moment. From a historical perspective (exhibited in Figure 2—2), we argue more new pricing models will be created often alongside with the innovative cloud technologies. We have observed many CSPs, such as Cloudheat [109], Databricks [111], Cisco systems and RingCentral [110] start to roll out a new pricing model that is supported by a hyper-converged solution to extend cloud computational power to the edge, which is close to the end-user. They call it as distributed or fog computing or data center in a box. This solution can eliminate network latency and routing path hops and provide much mobile computation power. Although this type of cloud service may still be in an incubation stage, they could become the major player. On the other side of the pricing spectrum (Refer to Figure 2—3),

other CSPs, such as Iex.ce [108], Cambridge Intelligence, Arkessa, and Vizolution extend cloud resource pools to a global market reach by leveraging blockchains and desktop grid technologies with a much competitive price for "Pay-per-Task" (See Figure 1—1) These practical cases illustrate that innovative cloud technologies with competitive new pricing models will stimulate new cloud service demands.

Practically, we can have at least 60 different pricing models for various cloud services. The reason to illustrate 60 pricing models is that different cloud services require a different approach to address various issues of cloud services, a method of delivery, payment, promotion, discrimination and etc. The detail of each pricing model is excluded from this chapter due to the limited space. The analysis results in this chapter for cloud pricing strategy are similar to Hinterhuber's findings shown in Figure 2—6, which the dominated pricing strategy is market-based pricing strategy (35 pricing models). Overall, this chapter has defined and highlighted many pricing model categories that have been already applied to different industries, especially service industries. Although many of them are not available in today's cloud market, CSPs should not eliminate their imagination to a few pricing models. As Weinman [1] indicated, CSPs should learn from other industries and compete on pricing, not on price alone. Table 2—4 provides the summary information of these categories of pricing models at a glance.

Table 2—4 Summary of Taxonomy Pricing Models

Name of the model category	Qty. of models	Sub-C Qty.	Simple Definition	Advantages	Disadvantages	Typical example of Applications
Service-Based	6		It is driven by the customer value proposition of "good to have" (select)	The value can be defined objectively	If the quantity grows fast, the cost could be out of control	SaaS delivery
Experience Based	4		The pricing category is driven by customers' value proposition of "good to do." It is equivalent to performance-based pricing	It is a win-win model and fair to both parties	Not every service can be specified with a list of performance metrics	On-line advertising camping
Customer value Based	4		The pricing models are driven by customers' value proposition of "good to be"	Maximized the sales profit-based customers W2P	Challenging to select the right service features for pricing models	Many services real including the real estate industry

Free Upfront and Pay Later	3		It is to leverage free products with minimum features for generating higher profits from premium customers	Increase customer base and market share	Challenging to decide product components between free and premium	E-commerce, pay-TV, proprietary software license
Auction	5		Price is settled by bidding-based rules in public	Price is transparent; Price is quick to be set down	The price is unpredictable	Real estate industry
Online	1		Price is published on a web page	No extra handling cost, Price is transparent	High risk of Security and privacy issues	e-commerce
Retail-Based (RB)	26		It is a B2C type of pricing model	The optimizing product set to maximize profit	Too many options	Retails industry or online retail
Sub-RB: Product Mix		6	Product-oriented pricing models	Boost sales, generate extra revenue	Lead to a bad reputation	Telco services
Sub-RB: Discounts		6	Payment driven pricing models	Increase cash flow	Reduce the profit margin	Nearly all retail industries
Sub-RB: Promotional		7	Sales strategy driver pricing models	Increase sales and reduce inventory stock	Reduce the overall profit margin	PC sales
Sub-RB: Discriminatory		7	Customer segmentation driver pricing models	Increase in profit margin	Increase in sales cost	Service industries and automobile retails
Expenditure-based	3		Price is decided by a proportion of cost or expenditure of production	More comfortable to be constructed and understood	Either overshoot or undershoot	Dominated firms often have a market monopoly
Resource-based	4		Price is decided by resources to provide the services	Consumers deploy scarce resource both efficiently and effectively	Providers have no incentive to optimize price	Professional Consulting industries
Utility-Based	4		Price is metered. Usage is monitored. Payment is according to usage or pre-defined plan in a regular term	Each customer can access the service that is unfordable by a single individual	Each individual has to rely on the utility service	Utility industries: gas, electricity, water supply, and sewage, telco
Total	60					

In the taxonomy of pricing models, this chapter emphasizes on value-based pricing strategy for cloud service because the nature characteristics cloud computing is service-oriented. However, it does not mean that cost-based pricing is not important. It often provides a bottom-line price for CSPs. The value-based pricing illustrates the maximum price, which is how much the cloud customers are willing to pay, while the market-based pricing will give CSPs an estimation of

competitive prices in the marketplace. If the cost-based pricing can set up the lower bound price, then the value-based price is to estimate the high bound. The market-based pricing gives a price variation between the lower and higher bounded prices. Cloud pricing strategies, tactics, and models are mainly dependent on various cloud services features, cloud technologies, targeted customers, market environment, cloud orchestration, and etc.

2.4. Survey of Pricing Models in Details

During the last decade or so, more than hundreds of papers are published regarding cloud pricing models. Many pricing models can be considered as an extension of the grid, cluster, distribution, high performance, parallel, Peer to Peer (P2P), network and utility computing. Based on the taxonomy criteria, the following survey will be organized into three cloud pricing strategies. This chapter selected published work between 2008 and present for in-depth diving investigation. One of the compelling reasons to select these research works is the majority of studies proposed either new mathematical solutions or novel ideas for various innovative pricing models.

According to the context of these papers and our criteria, we classify [49] [50] [52] [55] [56] [61] [62] as market-based pricing and [60] [78] [79] [80] [113] [114] [115] [116] [118] [119] as cost-based pricing, and [84] [85] [86] [91] [93] [95] and Chapter 3 as value-based pricing. This investigation highlights the uniqueness of their ideas, new concepts, and the contributions of each paper. Moreover, this chapter shows their relationship whether it is a continuation of previous work or the original work.

2.4.1. Pricing Models of Pre-Cloud Computing

In the later 1999 and early 2000s, Buyya et al. [49] [50] proposed a computational economy framework to regulate grid computing resources based on market supply and demand. The basic idea was to provide a set of different pricing models that can optimize grid resources and objective consumer functions through trading and broker services on an open commodity market. The authors introduced at least seven different types of pricing models: commodity market, posted price, bargaining, tendering/contract-net, auction, bid-based proportional resource sharing, community/coalition/bartering and monopoly & oligopoly models. In addition, the authors also indicated there were many challenges [51], such as managing grid resources, leveraging grid

technologies to allocate grid resource and implementing different pricing models. As a result, many proposed pricing models need further consolidation.

When virtualization has become a mature cloud technology during the 2000s, cloud computing was on the horizon. Based on many years' research experiences, Buyya et al. [52] [53] argued that the paradigm had shifted. The authors proposed the architecture solution for market-based pricing for cloud resource allocation. The solution was an extension of the grid computing [54]. The goal of this architecture is to create third-party services (or a cloud broker) to allow cloud consumers to utilize global cloud infrastructure effectively. The idea of global cloud or multi-cloud service providers was cutting edge at that time. It has only become practicable after the serverless container technology has emerged recently [108].

2.4.2. Market-Based Cloud Pricing

By leveraging Buyya's early proposal, Toosi et al. [55] developed a novel algorithm in combination with different cloud price models that a CSP can optimize its cloud capacity for cloud business revenue maximization. The main contributions of their research are: 1) present a stochastic dynamic programming technique to calculate the maximum number of reserved instances that a CSP can offer to cloud customer for its revenue maximization, 2) Due to the computational complexity of dynamic programming technique, the authors provided two heuristic algorithms. 3) The paper created a framework that is validated by a large-scale simulation dataset provided by Google. The following four equations can illustrate the essence of their solution shown in Figure 2—10

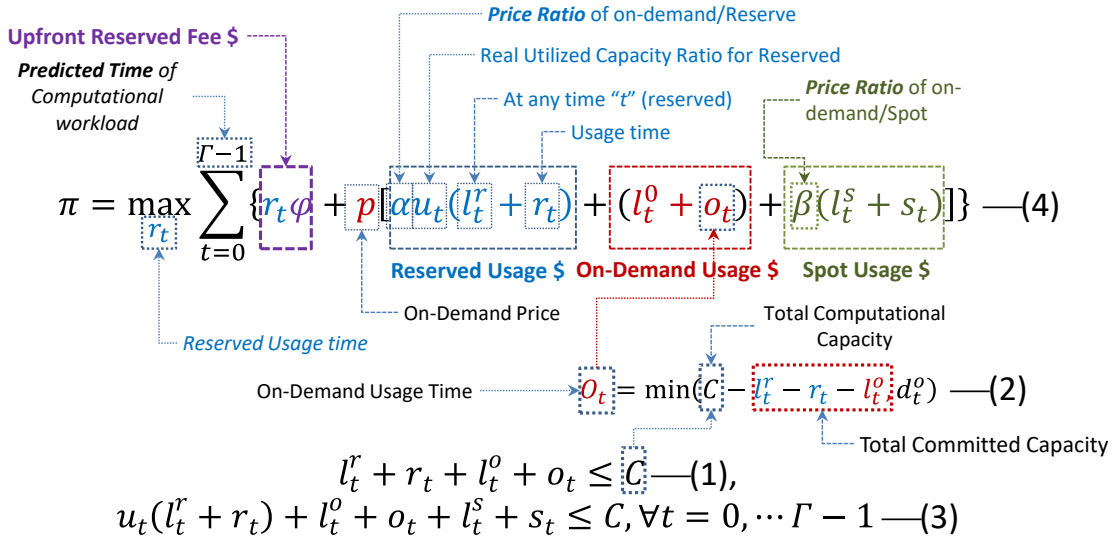


Figure 2—10 Optimization Solution for Cloud Business Revenue Maximization

Equations 1, 2, and 3 are three constraints. Equation 4 is the sum of quantity multiplied by unit prices all three revenue streams based on three price models: reserved, on-demand and spot. The paper presented a novel idea about how to maximize cloud revenue with a fixed cloud capacity. However, there are some gaps regarding pricing model assumptions: 1) the revenue function excluded the cost component; 2) AWS is charging on hourly base for on-demand instance while Google Cloud Platform (GCP) is charging on minute base; 3) Based on the AWS price model, spot instances can be terminated in 2 minutes warning advance. So, the l_t^s can be set to zero at any time and s_t can also be set to zero if there is an issue for cloud capacity contention. Overall, Toosi explained AWS's pricing models for a market-based pricing strategy by combining different pricing models to maximize CSP's revenue. The remaining challenge is how to model an arbitrary behavior of the instance termination.

Similarly, Xu et al. [56] tackled the same problem by introducing a dynamic pricing model that can be traced back to Gallego's work [57]. The main idea of their dynamic pricing model was to assume both arrivals $f(p)$ and departure $g[f(p)] = k[1 - (f(p)]$ (where, k >0) rates for AWS spot instance demand are a Poisson process. If the optimal stochastic policy changes price continuously (or the price change is a continuous variable), then the expected revenue function E_u and maximum profit $J^*(x,t)$ are shown in See Figure 2-11

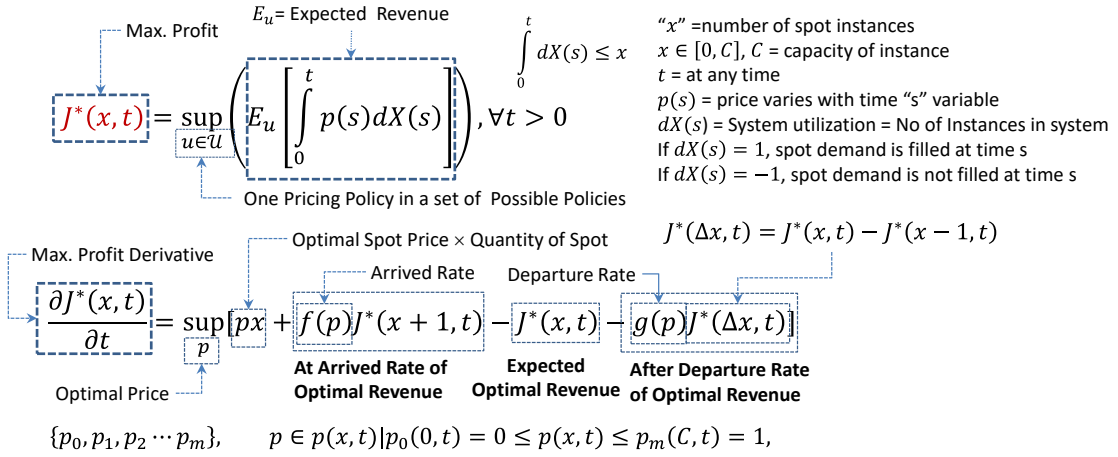


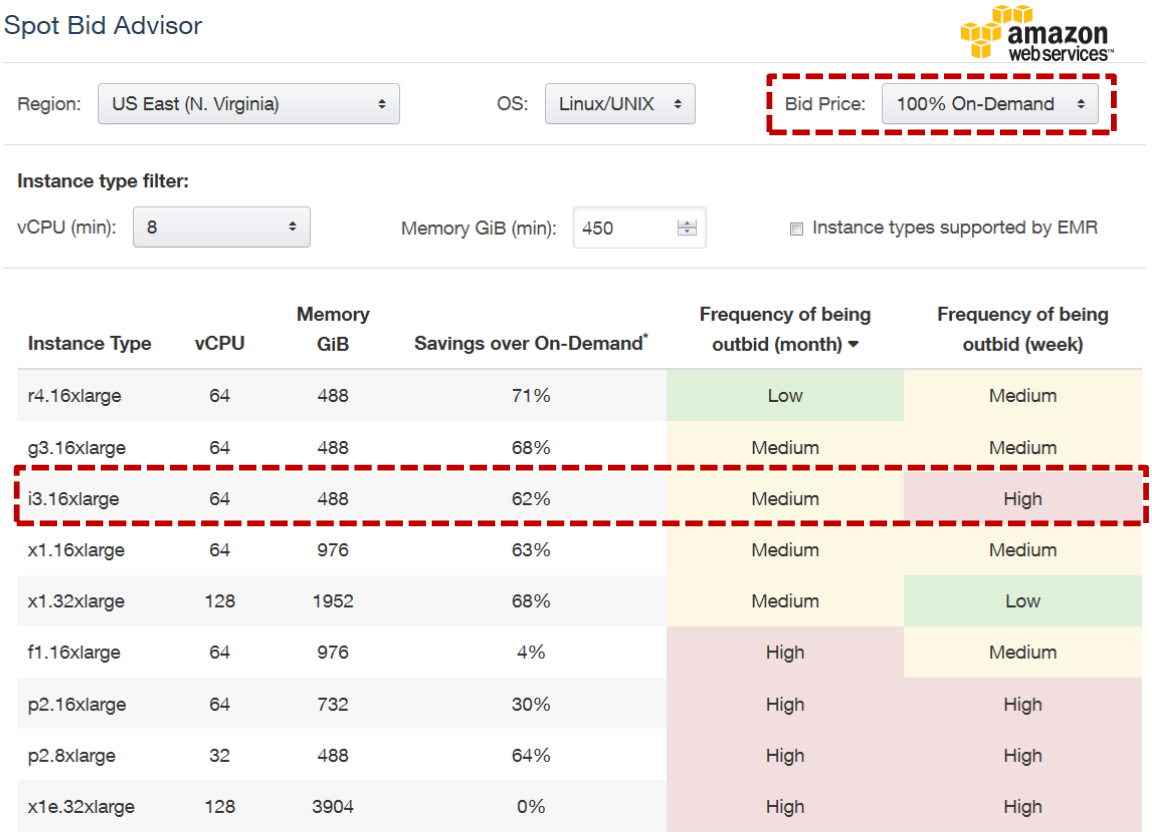
Figure 2—11 Dynamic Price Modeling AWS Spot Instance for Profit Maximization

The main contributions of this paper offer an alternative pricing model for CSP to price its spot instance dynamically. This means that a CSP reserves its right to change the spot price at any time. Moreover, this pricing model can provide a regulating tool for CSP to balance its limit cloud capacity resources and control or cap the spot instance demand and support on-demand and reserved instances. However, few assumptions need further consolidation:

The observation of spot price variation within a narrow band could be valid. It is right for a particular instance in the past. However, it is quite challenging to be generalized to all instances, zones, and regions in today's environment. Joshua Burgin (General Manager from AWS) indicated: "Prices for instances on the Spot Market are determined by supply and demand. A low price means that there is more capacity in the pool than demand. Consistently lower prices and lower- price variance mean that the pool is consistently underutilized. This is often the case for older generations of instances such as m1.small, c1.xlarge, and cc2.8xlarge." [58]. AWS "Spot Bid Advisor" shows many instances are frequently outbid shown in red in comparison to its on-demand price in Figure 2—12. In one case, the spot price reached a ridiculously high price - $999.00.[58]

Usually, the spot instance price variant with time is neither convex nor continuous. As Gallego [57] noticed that "the stochastic optimal policy changes prices continuously and thus may be undesirable in practice" Both arrival and departure functions are defined as more like a power function rather than a Poisson distribution function because (see Equation 2-1)

$$f(p) = k(1 - p^a)^b, g[f(p)] = k[1 - f(p)] \quad (\text{where, k} > 0, \text{a} > 1, 0 < \text{b} < 1) \tag{2-1}$$



Spot Bid Advisor

amazon web services

Region: US East (N. Virginia) OS: Linux/UNIX Bid Price: 100% On-Demand

Instance type filter:

vCPU (min): 8 Memory GiB (min): 450 Instance types supported by EMR

Instance Type	vCPU	Memory GiB	Savings over On-Demand*	Frequency of being outbid (month)	Frequency of being outbid (week)
r4.16xlarge	64	488	71%	Low	Medium
g3.16xlarge	64	488	68%	Medium	Medium
i3.16xlarge	64	488	62%	Medium	High
x1.16xlarge	64	976	63%	Medium	Medium
x1.32xlarge	128	1952	68%	Medium	Low
f1.16xlarge	64	976	4%	High	Medium
p2.16xlarge	64	732	30%	High	High
p2.8xlarge	32	488	64%	High	High
x1e.32xlarge	128	3904	0%	High	High

Figure 2—12 AWS Spot Bid Advisor

The model also excluded the cost component for CSP's revenue maximization based on Greenberg [60] works. This assumption could be inaccurate to interpret Greenberg's work. The paper also assumed that cloud customers are price takers. AWS has full control of the spot instance based on both arrival and departure rates. It means that AWS has control of the spot instance's bidding process.

So, the question is how the AWS controls or regulates its spot instance and what mechanism is behind the AWS' spot instance bidding processing. Before our further discussion of the AWS spot instance, it is important to understand how it works. AWS spot instance bidding mechanism is similar to the first price sealed-bid (FPSBA). It is a prevalent auction practice in the real estate industry, which is called "Sale by Set Date." In contrast to the English auction process, it is a blind auction, which all the bidders submit their bidding prices simultaneously without any pre-

knowledge of other bidding prices — the highest price the bidder wins the cloud instance time slot. However, the price that the highest bidder pays for is the market price, not his bidding or reserved price. For example, the highest bidder's reserved price is $2.00, but the next highest bidding price is only $1.00, the highest bidder only pays $1.01, not $2.00.

AWS might have its own reserved price with different types of spot instances cross different regions and zones based on the availability of its cloud infrastructure capacity after satisfying its "on-demand" and reserved customers. When a new bidder submits a fresh bidding price that is higher than the old bidder's reserved price at any time, the old bidders have two minutes of warning time to terminate their running instances. In this case, AWS will not charge its customers if the instance running time is less than one hour. The existing customers can either revise their upper ceiling reserved price or move their workload to "on-demand" instances. As we illustrated above, the bidding price might be well above the "on-demand" price. It might sound irrational, but if a customer only pays a very short period. It will become acceptable if the average price is still less than the "on-demand" price. Recently, AWS has capped four times of "on-demand" price as the highest bidding price. Moreover, AWS also offers up to 6 hours of spot instances (Spot block in 2015) to accommodate different types of workloads. These new rules will change the bidding game.

Orna et al. [61] provided a different interpretation of AWS spot pricing, which they show a mechanism of AWS spot instance via a reversed engineering based on the traceable data or files (from Tim Lossen's Cloud Exchange and Kurt Vanmechelen's Spot Watch) in Apr. 2011. They concluded that AWS sets its spot instance price in a random auto-regression manner. For the high bound price, it is set to reflect a market-driven mechanism. For the lower bound, it is reserved within a narrow band, which shows as Equation 2-2:

$$\delta_i = -a_1\delta_{i-1} + \varepsilon(\sigma), \ \ and \ \ p_i = p_{i-1} + \delta_i \qquad (2\text{-}2)$$

where, $\delta_i$is the narrow band, $a_1$is the coefficient, $\varepsilon(\sigma)$ is the white noise, p_i is a price at any time "i." It is an empirical observation. The goal of the paper was to help cloud customers to understand AWS spot mechanism in order to bid the spot price.

To answer similar questions, Zheng et al. [62] presented spot price bidding models or strategies for different types of workloads. The authors' conclusions are their bidding strategy can reduce 90% of the cost in comparison with the "on-demand" price. The paper assumed two types of scenarios, which are one-time bidding and continuous bidding strategies. For the one-time bidding strategy, the cloud consumers can achieve the lowest possible bid price p^* illustrated as following Figure 2—13.

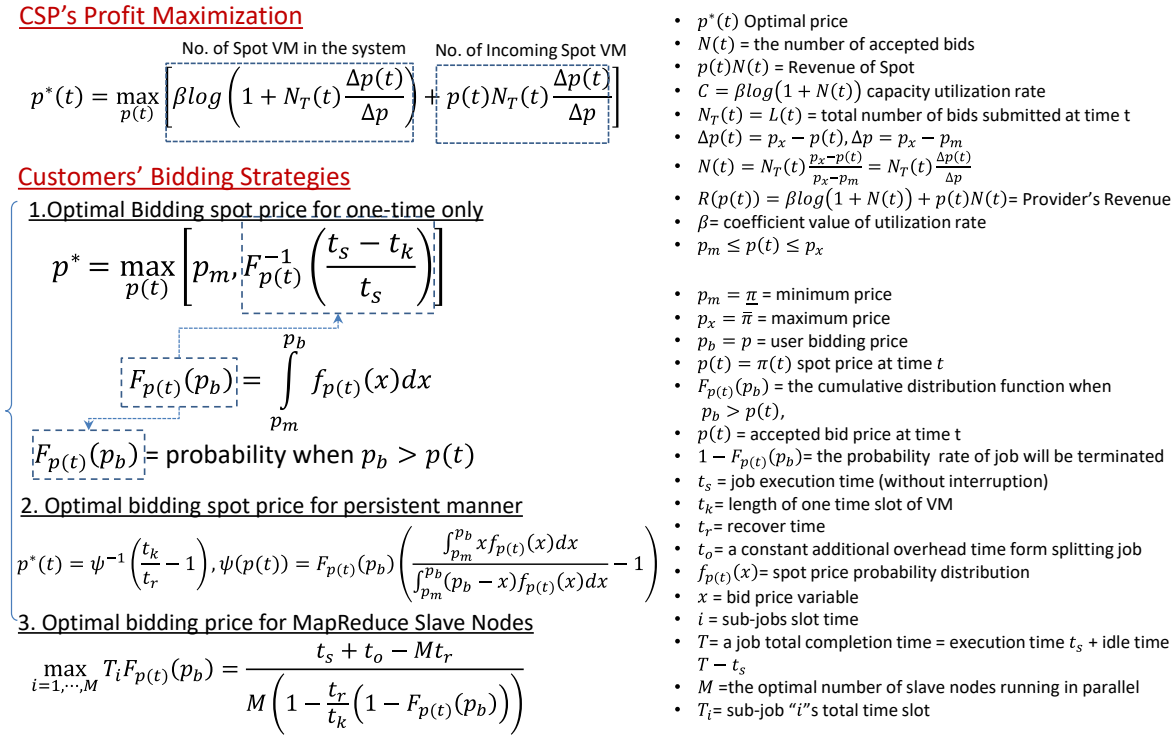


Figure 2—13 CSP's Profit Maximization and Customers' Bidding Strategies to Minimization Spot Price

Zheng's work can be summarized into three main contributions of AWS spot instance pricing bid strategy: 1.) Price orientation bid strategy, 2.) SLA priority bid strategy, and 3.) MapReduce workload application. Based on the authors' observation, they conjecture that only a few users bid for spot instances due to heavy-tailed spot price distribution. However, the gaps in the paper are: 1.) they assumed that the highest spot bid price should be less than the on-demand price, but the reality is the bid price could exceed the on-demand price. 2.) The maximum revenue function analysis did not include the marginal cost from a CSP perspective. 3.) The authors did not give a further explanation of the capacity utilization function, 4.) The assumption of uniform distribution for bid prices appears to be contradicting the following contents of the bid price distribution,

namely Pareto and exponential distribution. 5.) The paper intended to isolate the issue of the spot resource from other on-demand and reserved resources, but in reality, the CSP has a big cloud resource pool for all price models. 6.) The assumption of workload is i.i.d needs further clarification.

Overall, the possible spot pricing model serves well for interruptible workloads. These jobs have some essential characteristics 1.) Running time for the job is unpredictable, 2.) It has many checkpoints 3.) The job can continue to run after any stop point, 4.) It works well for stateless [11] applications (The server does not save the client's data that is generated in one session). Based on the paper's final discussion and conclusion, the spot pricing bid strategies are only applied for interruptible workloads rather than all types.

Since AWS launched its spot instance in 2009, it has generated enormous interest in the academic world. The amount of published papers [63] [64] [65] [66] [67] [68] [69] [70] [72] regarding of AWS spot pricing model is overwhelming. Perhaps, it has a large price discount in comparison with "on-demand" and reserved price models. The basic idea of a spot instance mechanism can be considered as an analogy of a spot price of electricity in an energy market [71]. Most of SLA and cost-oriented papers presented some impressive and complicated mathematical formulas based on both historical spot price data and subjective assumptions. However, AWS can terminate any spot instance at any time, although it gives you only 2 minutes of advance warning time. It is very challenging to consider any logic or rational pattern behind AWS to terminate any spot instance.

A SaaS company –MOZ's experience in 26/Sep/2011 [59] provided a perfect example shown that it would be a very high risk to rely on the spot instance price alone for SLA services delivery. Due to MOZ out of the bid [12], all MOZ [13] services had been shot down [73] . It took MOZ 14 days to restore its services fully. MOZ has about 26,474 subscribers plus 5,000 free trial customers. If we assume MOZ's customers pay premium $599/ per month, the estimated revenue loss is about $8 million in 14 days if we do not take consideration of potential new incoming subscribers,

[11] Statefulness means a backend hosting server or VM maintains user's state information in the sessions form. In contrast, Stateless does not keep any state information for the end-user. Anything is stored on the end-user or client's side in the form of a cache.
[12] MOZ reserved bid was $2/per instance for more than 3 years
[13] MOZ provides Search Engine Optimization (SEO) web crawler services to its customers. MOZ charges its customers on monthly subscription fee.

customer experiences and the company's brand and reputation. That is why MOZ had switched its cloud infrastructure from a public cloud to colocation [74] in 2013

Usually, the spot pricing instance is not suitable for mission-critical applications, but many large and medium-sized enterprises or even some small firms require mission-critical infrastructure. Spot instances could be applied to interruptible workloads. However, some computation-intensive workloads, such as batch processing, encoding or decoding, rendering, modeling or continuous integration, cannot generate checkpoints over its multi-hour running period. Other leading CSPs do not offer spot pricing models, but they provide a fixed discount price with limited cloud service features, which are similar to AWS' spot instance. It means there is no free lunch.

2.4.3. Cost-Based Cloud Pricing

As early as 2008, Greenberg et al. [60] discussed the cost-based strategy regarding cloud data centers. It provided a rough estimation of infrastructure costs for cloud services. Some critical assumptions of their estimation were 50,000 physic servers or nodes and 5% of an interest rate for capital investment, $3,000 per server, a three-year lifecycle time and an electricity price of $0.07/per Kilowatts hour (KWH). The guideline to build its own cloud data center showed in Table 2—5

Table 2—5 Cost Guideline of Cloud Data Center

Amortized Cost	OECD Electricity price in 2014	Cost Components	Sub-components
~45%	~37%	Servers	CPU, RAM, Storage Systems
~25%	~20%	Infrastructure	Power distribution and Cooling
~15%	~30%	Power draw	Electrical Utility Costs
~15%	~12%	Network	Links Transit Equipment

The authors highlight major issues across many data centers at that time (before 2008), which has a lower utilization rate of data center resources. They identified some approaches to increase the data center efficiency, such as optimize the data center internal network, design market-based algorithms for data center utilization and improve inter-connected data center networks. We argue the estimated costs for the cloud data center are dependent on each case and the location of a data center. For example, the authors assumed the electricity price is $0.07/per KWH. This price estimation is on the lower end [75]. The average price of electricity power cross developed nations (OECD) is US$0.23 [76]. Even in the US, the average price of household electricity is around 0.125, and the industrial price is about $0.10. If we use OECD average price and keep other cost

items unchanged, the proportion of each cost component for the amortized cost will be changed dramatically. The portion of the amortized cost of electricity will be double. Moreover, the paper did not include the data center space cost, which is another significant cost item. It could be up to 15% [75] of the total cost of a typical cloud data center. Nevertheless, the paper made a significant contribution to cloud data center price estimation. They are the pioneer of cost-based pricing.

In comparison with Greenberg's approximation estimation, Walker [77] [78] laid out the precise costs of both CPU and storage for Net Present Value (NPV) in comparison with AWS EC2 and S3 (or public cloud) presented in Figure 2—14

$$\Delta \boldsymbol{NPV} = \sum_{T=0}^{N} \frac{C_T - E_T + L_T}{(1+I_F)^T} + \frac{S}{(1+I_F)^N} - C, \quad \boldsymbol{S = \gamma \times \Omega \times [V_T]_{\Omega} \times K \times e^{-0.438T}}$$

$$\boldsymbol{C_T} = -\rho \times H_T - (365 \times 24) \times \delta \times (P_C + P_D \times [V_T]_{\Omega})$$

$$\boldsymbol{E_T} = (1.03 \times [V_T]_{\Omega} - [V_{T-1}]_{\Omega}) \times \Omega \times \mathrm{K} \times e^{-0.438T}$$

- δ: Cost of electric utility ($/kilowatt hour)
- Ω: Size of purchased disk drives (Gbytes)
- ρ: Proportional difference between human effort in maintaining a purchased versus a leased storage infrastructure
- $\boldsymbol{\gamma}$: Used disk depreciation factor on salvage ([0.0, 1.0])
- C: Disk controller unit cost ($)
- H_T: Annual human operator salary ($)
- I_F: Risk-free interest rate (%)
- K:Current per-Gbyte storage price ($/Gbyte)
- L_T: Expected annual per-Gbyte lease payment ($/Gbyte/year)
- P_C:Disk controller power consumption (kW)
- P_D:Disk drive power consumption (kW)
- V_T:Expected storage requirement in year T (Gbytes)
- $\Delta \boldsymbol{NPV}$:Incremental Net Present Value
- $\boldsymbol{C_T}$:the operating cost in year T,
- $\boldsymbol{E_T}$:the capital cost in year T
- S:expected end-of-life disk salvage value

Figure 2—14 NPV Value Versus Leasing Public Cloud

According to Walker's calculations with assumptions of 90% of server utilization rate, 5% of a capital cost, and clusters of 60,000 CPU cores capacity, he concluded that a three-year investment commitment is the optimal term length for purchase case because of the lowest cost per CPU hour. Second, the operational lifespan should be within ten years. Moreover, if the lifespan is less than two years, it would be cheaper to lease computational capacity (off-premise). Finally, if the capacity utilization rate is less than 40%, it would always be more reasonable to use cloud resources (off-premise). Based on the same principle of NPV, Walker demonstrated formula for the enterprise storage cost in the comparison between own build (on-premise) or purchases and public cloud (off-premise) shown in Figure 2—15

$$NPV = \sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T}, \qquad NPC \times \sum_{T=0}^{Y-1} \left(\frac{1}{\sqrt{2}}\right)^T \Rightarrow NPC = TC \times \frac{1-\left(\frac{1}{\sqrt{2}}\right)^Y}{1-\frac{1}{\sqrt{2}}}$$

$$NPC = Y \times TC, \qquad TC = TCPU \times H \times \mu, \qquad R = \frac{NPV}{NPC}, \qquad PC = \frac{FC}{\left(\sqrt{2}\right)^T},$$

$$R_p(purchase) = \frac{NPV}{NPC} = \frac{\left(1-\frac{1}{\sqrt{2}}\right) \times \sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T}}{\left(1-\left(\frac{1}{\sqrt{2}}\right)^Y\right) \times TC}$$

$$R_{up}(purchase-upgrade) = \frac{C_0 \times \sum_{T=0}^{Y-1} \frac{C_T - A}{(1+k)^T}}{Y \times TC}, \qquad Versus \quad R_l(lease) = \frac{\sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T}}{Y \times TC}$$

$\boldsymbol{R_p}$: The real cost of CPU hour for purchase case
Y: Lifespan of Year
K : Cost of capital (interest rate)
T: Asset value at Year T
C_0: Hardware acquisition cost
C_T: Operation cost at each year T
TC: Total Useful Capacity
A: The server cluster's original purchase cost

R_{up}: The real cost of purchase and upgrade case
R_l: The real cost of Lease
PC: Present Capacity
FC: Future Capacity
NPC: Net Present Capacity
μ: The expected server utilization rate
TCPU: Total CPU cores the server cluster
H: The expected number of operation hours

Figure 2—15 Storage Pricing Comparison between Purchasing and Leasing

The hypothetical assumption for cloud storage pricing was based on the threshold levels of storage illustrated in Table 2—6. It means that CSP often gives a volume discount, which is a kind of linear discount rate.

Table 2—6 Hypothetical Assumption of Cloud Storage Pricing Structure (2010)[78]

Default	Storage > 50 TBytes	Storage > 100 TBytes	Storage > 500 TBytes
$0.15/Gbyte/month	$0.14/Gbyte/month	$0.13/Gbyte/month	$0.12/Gbyte/month

However, the reality is that the storage price is quite challenging to be generalized because each CSP will have a different cost-based pricing model for cloud storage (as shown in Table 2—7). The price range could be as high as 21 times difference, which is dependent on many storage performance factors. Moreover, each CSP may give different depreciation rates of cloud storage prices each year. This means the L_T (Expected annual per GB lease payment) is a time variable, not a constant.

Table 2—7 Cloud Storage Pricing From Different CSPs (in 2017) [58] [75]

Cloud Service Provider	Storage ($/GB/Month)	Download ($/GB)
Backblaze	$0.005	$0.02
AWS S3	$0.021	$0.05
Microsoft Azure	$0.022	$0.05
Google Cloud Platform	$0.026	$0.08
Softlayer	$0.10	$0.09
Rackspace	$0.105	$0.12

Walker's suggestion was if a decision-maker wants to have cloud storage resource for more than four years, the solution of building own storage infrastructure (on-premise) is a preferred option otherwise cloud solution (off-premise) would become a favorite option because of a higher NPV value. The main contribution of Walker's papers is it demonstrated how to use the NPV to construct a cloud cost-based model by taking consideration of Moore's law or IT assets depreciation within a specified period. However, the predicted cost per Gbytes is dependent on previous observation. Different sources of price data collection could lead to different results. For example, if we adopt McCallum's dataset [79], the $G_x = 1.3314e^{-0.06T}$ (the depration rate of $/per GB) between Apr-2003 and Sep-08 (Refer to Figure 2— 16 a) Moreover, if we take the period from 2003 to 2017, the best format to fit the historical HDD price data set would be logarithm rather than an exponential one (Refer to Figure 2— 16 b). $G_x = -0.306\ln(T) + 1.3466$. The R-square value is 0.8925. Finally, if we take the time span from 2008 to 2017 and change the price scale from dollar /GB to dollar /TB, the coefficient of the fit equation would change again: $G_x = -41.3\ln(T) + 196.83$. The R- square value is 0.9183 (Refer to Figure 2— 16, c)

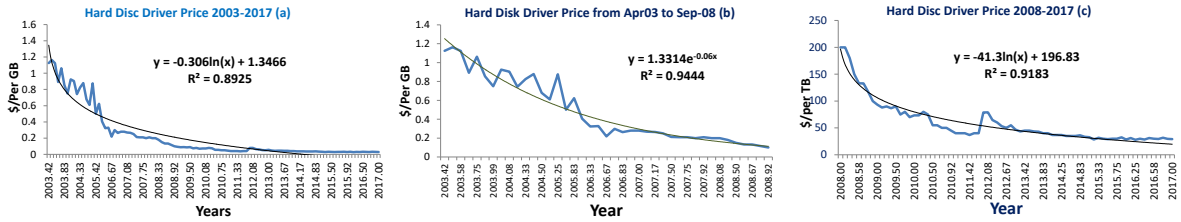


Figure 2—16 Hard Disk Drive Price

It indicates that E_T (a capital cost in year T in Figure 2— 14) is dependent on the number of observation years (or data points) and the unit of time span and unit price/per HDD. If these variables are changed, the fit-equation and its coefficients will also be changed. Subsequently, the decision model is oscillating according to different time spans. Walker's cost-based pricing can be considered as a root of resource performance driven by cloud customer's NPV. If we shift our

focus from cloud customer to CSP, a value proposition becomes an issue on how to optimize the finite capacity of the cloud resource pool

Xu et al. [80] proposed a preliminary price model for cloud resources. The basic idea of their model is derived from the alpha-fair utility function as an economic utility function, which is the same as the Isoelastic or constant elasticity function (a particular case = constant relative risk aversion (CRRA) of Hyperbolic Absolute Risk Aversion (HARA)) based on economic utility theory [81]. It means a CSP seeks to maximize its revenue if cloud consumers make rational choices with risk aversion preference. If the CSP wants to maximize its revenue, it could have five different strategic options for pricing: (1) basic, (2) the 1st order price discrimination, (3) throttling, (4) SLA performance, and (5) profit maximization illustrated in Figure 2— 17 and Figure 2— 18, Equations 6 and 7 of Figure 2— 18 shows how to maximize CSP's reverence with a capacity constrained.

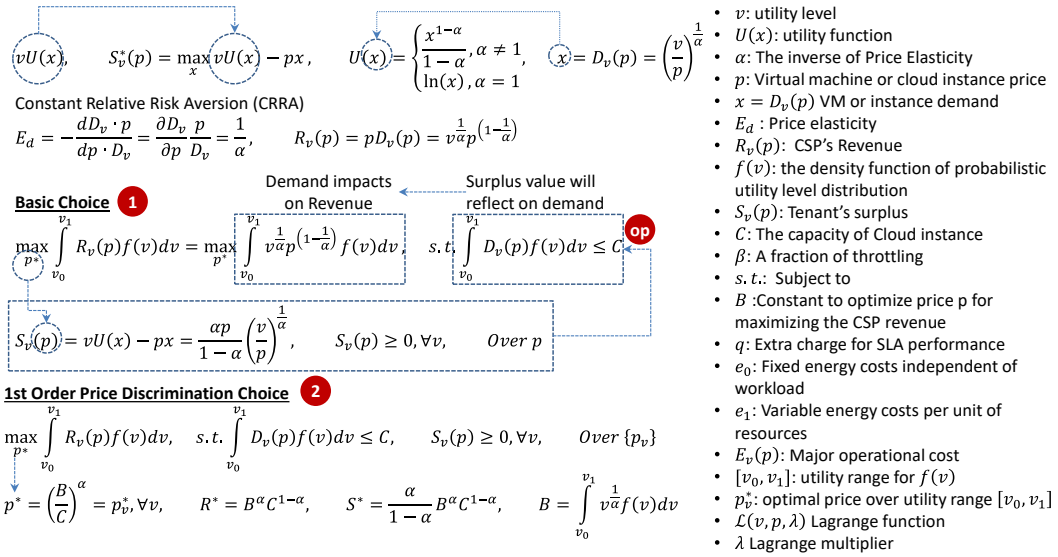


Figure 2—17 CSP's Revenue Max. Basic, the 1st Order Price Discrimination

The main contribution of Xu's paper is that it articulated various CSP's pricing choices by exploring the iso-elasticity function as a cloud customer's utility. The author demonstrated that CSP could leverage customers' surplus values to maximize its revenue if there is only one type of utility. However, there are a few practical gaps: 1.) the customer utility function and alpha-fair utility are two different concepts. One is the utilization rate of the limited amount of cloud capacity, and others have an economic connotation, which is to measure customer's subjective experiences. If a preference can be measured by a marginal value, it is a cardinal utility. Otherwise,

it will be measured by a ranking order, which is an ordinal utility. 2.) As authors indicated in their paper, it is challenging to charge cloud consumers with 1st order degree price discrimination because of the price transparency. In practice, it is more likely to adopt the 2nd degree (volume discount) and the 3rd degree (different prices to different consumer groups) pricing discrimination. 3.) The assumption of throttling requires further consolidation because the characteristics of online pricing, CSP has to declare its performance of cloud resources. If a CSP reduces the specified VM performance (such as CPU speed, RAM, and storage size), this means it cannot fulfill its legal obligation. An alternative option is to declare the cloud performance in a rough range. For example, AWS specifies its network performance as low, low to moderate, moderate, high. AWS does not provide a quantitive specification. 4.) It is not practical to assume that all cloud consumers have the same utility functions. 5.) A probability density function $f(v)$ needs further clarification. In addition, Google Cloud Platform (GCP) and the Amazon Web Service (AWS) pricing models are different (Refer to Table 2—2)

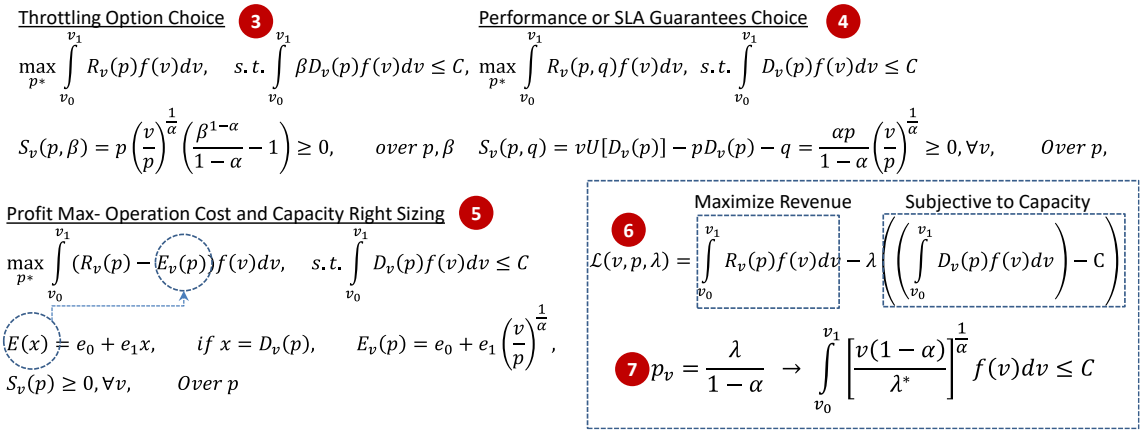


Figure 2—18 CSP's Revenue Max. Strategies for Throttling, SLA and Profit Max

Furthermore, the assumption of elasticity $E_d = \frac{1}{\alpha} = 3$ should require a further explanation because this parameter will impact on the shape of the utility function, which ultimately will determine the optimal price. Subsequently, the level of utility $v = p\sqrt[3]{x}$ If we use the paper's price assumption: p= 0.08/per hour for a small Linux instance, then utility level $v = 0.08\sqrt[3]{x}$. And then, the paper used Google, RICC and ANL cluster trace information to validate the utility density distribution. Based on the Alam et al. [84] research work, the workload pattern of Google cluster trace is more like the trimodal pattern rather than a convex. In addition, RICC is a parallel

computing cluster [83], and ANL is a grid computing cluster [84]. It would be very challenging to use these datasets for validation purposes of cloud resources modeling.

Although the paper had included a cost component in the equation of profit maximization, it excluded this critical element from other inference. Practically, the revenue maximization is not equal to profit maximization. Sometime, it might mean losing money if the marginal cost is higher than the sales price, which the higher revenue, the larger deficit is. According to Belleflamme and Pietz [85], the above revenue maximization function (monopoly pricing formula) should be altered as (Equation 2-3)

$$\max_{D_v(p)} \pi(D_v(p)) = D_v(p)p\big(D_v(p)\big) - C(D_v(p)) \tag{2-3}$$

where $C(\cdot)$ is an average cost and both price p and cost $C(\cdot)$ are the functions of demand: D_v, and demand is a function of p. Conversely, the price is also a function of demand: $p = D_v^{-1}(p)$. It would be a challenge to find an optimal value of p.

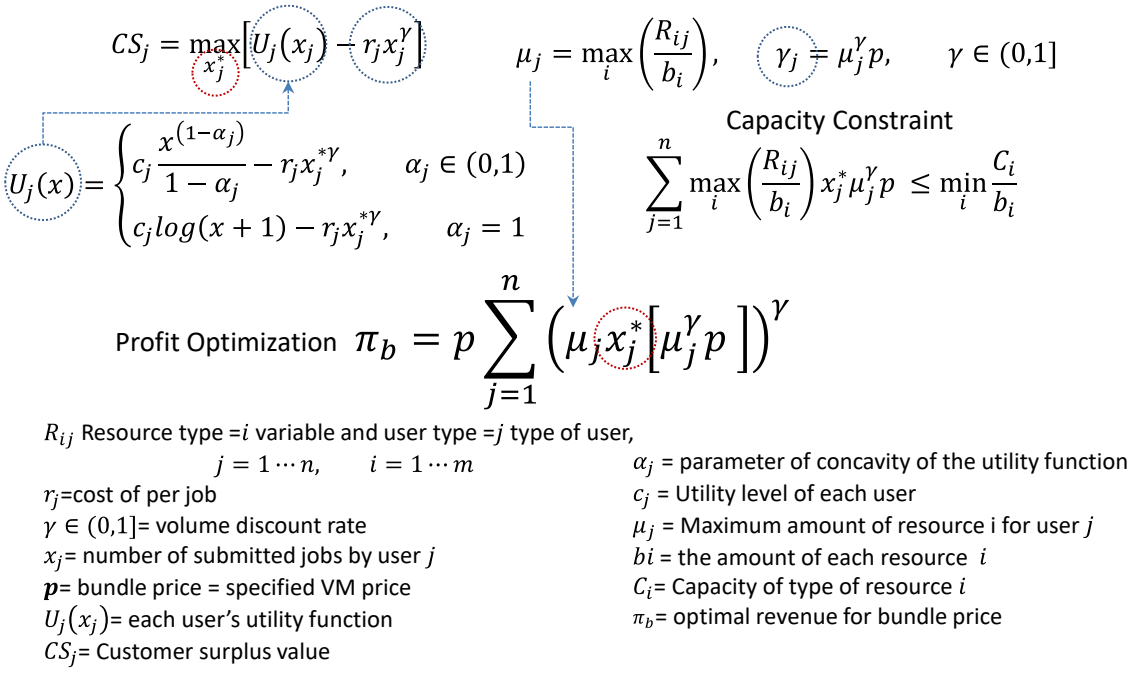
If we trace the root of Xu's research work, we can find Xu's cloud pricing model can be considered as an extension of Joe-Wong and Sen's [113][114] work. The difference was that Xu introduced a probability density function for cloud market demand. Joe-Wong and Sen proposed an analytical or mathematical framework of cloud pricing to optimize resource allocation, fairness, and revenue with a finite capacity of cloud resources. The core idea of their pricing model can be further traced back to Chiang et al.'s [115][116] study of network utility maximization (NUM). The essence of Joe-Wong's work can be summarized in the following mathematical pricing models shown in Figure 2— 19

As authors have noticed that "the function of π_b is a non-differentiable function of the amount of each resource i (e.g.b_i)." Subsequently, the value of b_i is a constant. This result actually reflects on a common practice in the cloud industry that was summarized by Kilcioglu and Rao [117], which any price of AWS MV can be presented as a proportion to the price of a base unit of VM configuration. Mathematically, Equation 2-4 shows this relationship. In other words, b_i is equal to 2^{k-1} for the majority of AWS VMs.

$$p_k = 2^{k-1}p_0 \tag{2-4}$$

where p_0 is the price of the smallest VM size, p_k is the k size of VM and $k = 1,2,\cdots$ is the number of VM sizes offered by a CSP. The distinct advantage of adopting this price model is that

the CSP can build a large VM resource pool at the finest granular level of scalability for cloud capacity and minimize a footprint of cloud infrastructure in a cloud data center.



R_{ij} Resource type =i variable and user type =j type of user,
$j = 1 \cdots n, \quad i = 1 \cdots m$

r_j=cost of per job
$\gamma \in (0,1]$= volume discount rate
x_j= number of submitted jobs by user j
$\boldsymbol{p}$= bundle price = specified VM price
$U_j(x_j)$= each user's utility function
CS_j= Customer surplus value

α_j = parameter of concavity of the utility function
c_j = Utility level of each user
μ_j = Maximum amount of resource i for user j
bi = the amount of each resource i
C_i= Capacity of type of resource i
π_b= optimal revenue for bundle price

Figure 2—19 Profit Optimization for Fixed Configuration of VM instance

By a similar line of reasoning for the network-oriented cloud price modeling, Shahrad [118] proposed a novel price model so-called Graceful Degradation (GD) to increase its cloud business profit by improving its cloud infrastructure (data center capacity) utilization rate and efficiency. The key idea of the GD pricing model is a self-capping mechanism, which is to "absorb demand fluctuation and reduce spare capacity." In other words, the GD price model is a cloud capacity regulator to smooth Service Providers' (SP, or business customers) demand between peak and valley. Their pricing model was built upon a function that is similar to the Cobb-Douglas utility function (Equation 2 in Figure 2—20) for an SP revenue function, which was reduced to the alpha-fair function (Item 3 Figure 2—20) regarding the total deliverable capacity and service degradation factor.

The significant contribution of Shahrad et al. work was the novel idea of leveraging a fine-grain pricing model to the regulator, a CSP's limited cloud capacity, which is a hybrid pricing solution to balance customers' demand and limited cloud capacity by brownout mechanisms (similar to electricity supply). The aim of this pricing model is to find a win-win solution for both customers

(gain price discount) and CSP (improve cloud infrastructure utilization rate). Later, Shahrad et al. [119] applied the same principle for customers' SLA delivery. In comparison with many previous works, they included a cost component in a profit maximization function shown in equation 1. To achieve the optimal value of c_b, the profit function $E(p)$ is selected to be differentiable.

$$R(c,\theta) = \gamma \theta^k c^\beta, \quad (2) \quad k = \beta = 1-\alpha, \qquad \alpha \in [0,1)$$

$$\gamma = \frac{1}{1-\alpha}, \qquad c_d = \theta c, \qquad \theta = \frac{c_d}{c}, \qquad (c_d < c), \qquad \theta \in [0,1]$$

$$\theta = 1, \qquad c_d > c_{max}$$

$$R(c,\theta) = \frac{(\theta c)^{(1-\alpha)}}{1-\alpha} = \frac{c_d^{(1-\alpha)}}{1-\alpha} \quad (3) \quad R(c,\lambda\theta) = \lambda^{(1-\alpha)} R(c,\theta)$$

$$R(c,\theta_1) \geq R(c,\theta_2), \qquad if\ \theta_1 \geq \theta_2, \qquad R(c_1,\theta) \geq R(c_2,\theta), \qquad if\ c_1 \geq c_2$$

$$\boldsymbol{E(p) = E(R) - E(Y)} \quad (1)$$

Revenue of Cloud Generating $= \int_{c_{min}}^{c_d} R(c,1) f(c) dc + \int_{c_d}^{c_{max}} R\left(c, \theta = \frac{c_d}{c}\right) f(c) dc$

Cost of Cloud $- \left(p_b c_b + \int_{b}^{c_d} p_d (c - c_b) f(c) dc + \int_{c_d}^{c_{max}} p_d (c_d - c_b) f(c) dc \right)$

$$\frac{\partial E(p)}{\partial c_b} = 0, \begin{cases} c_d \geq c_b \geq c_{max} & \text{Deliverable} \geq \text{Reserve} \geq \text{Max} \\ c_d \geq c_{max} \geq c_b & \text{Deliverable} \geq \text{Max} \geq \text{Reserve} \\ c_{max} \geq c_d \geq c_b & \text{Max} \geq \text{Deliverable} \geq \text{Reserve} \end{cases} \rightarrow \boldsymbol{c_b^*} \text{ and } \boldsymbol{c_d^*}$$

- p_b= Reserved capacity unit price
- p_d= On-demand capacity unit price
- C= Total Capacity of an Cloud Service Provider
- c_b= reserve capacity
- c_b^* = optimal c_b
- c_d= total deliverable capacity
- c_d^* = optimal c_d
- c_{min} = Minimum capacity demand in a period
- c_{max} = Maximum capacity demand in a period
- c= real time usage capacity
- θ= Service Degradation Factor
- $R(c,\theta)$= An SP's revenue function
- c = An SP's aggregate capacity demand
- k = Degree of homogeneity for revenue function
- $f(c)$ = Capacity distribution function in a period
- α = alpha-fair coefficient
- $E(Y)$= Expected payment of SP
- $E(R)$= Expected Revenue of SP
- $E(p)$= Expected Profit of SP

Figure 2—20 Pricing Model for Business Customer to Self-Cap its Cloud Capacity

2.4.4. Value-Based Cloud Pricing Strategy

For the value-based cloud pricing, one of the scientific approaches is known as a hedonic model. It has been widely applied to the consumer price index (CPI) by many OECD countries, such as Australia Bureau of Statistics (ABS), US Bureau of Labor Statistics (BLS), British Office for National Statistics (ONS), Germany Federal Statistical Office (Destatis), etc.

El Kihal et al. [84], Weinman [1], Mitropoulou [85] and Zhang [86] either proposed or presented a hedonic pricing model for cloud services. El Kihal showed the comparison results among major

CSPs (AWS, IBM Cloud, Microsoft Azure, Terremark, and Google App Engine) in terms of three cloud characteristics: memory ($ per GB), CPU ($ per CPU) and Storage ($ per 100GB). The hedonic function is shown in Figure 2—21. Overall, the paper had a gap to explain the details of how the dataset was collected and how many cloud instances were gathered.

The experiment result indicated that an adjusted R-squared value of the linear regression was between 0.43 and 0.69 (or 0.76 for Terremark). The interpretation of their experiment results seems to be unclear. Ideally, the constant coefficient of linear regression should be equal to zero because none would like to pay the monthly fee for no hedonic characteristics (RAM =0, CPU=0, and Storage =0). If the constant is not equal to zero, it often means a fixed effect. Otherwise, the linear regression model has some issues. Checking the adjusted R square values, it only explained 43% ~ 69% of the data. Both IBM and Microsoft's adjusted R square values were less than or equal to 50%. It might indicate the linear equation is not "goodness of fit."

$$BA_p = \beta_{0p} + \sum_{i \in I} \beta_{ip} x_i + \varepsilon_p \, , [p \in P]$$

BA_p: Price plan or billing amount
P: Number of CSP
β_{0p}: the constant coefficient of linear regression
β_{ip}: Parameters of hedonic characteristics
x_i Hedonic Characteristics
i: number of hedonic characteristics
ε_p: Error term of the regression equation
I: Total number of Characteristics
P: Total number of CSP

Figure 2—21 Hedonic Pricing Model for Cloud Services

In comparison to El Kihal et al. [84] paper, Mitropoulou et al. [85] made some progress of the hedonic method. Their work explained how and where the dataset was collected, but the author did not generalize the hedonic linear equation. Moreover, the adjusted R2 value of the experiment is only 57.5% and 53.7% for linear and exponential models, respectively. It means the linear model can just explain 1,577 out of the total of 2,742 data points. Nevertheless, the paper added three more cloud characteristics (RAM, CPU, Storage, OS, Transfer-Out and Subscription) for the hedonic calculation. The primary issue of the paper is that if the paper adopted a hedonic index measurement, it needs a base period for comparison.

This issue was solved by Zhang's works [86] based on Pakes [87]'s seminal work. The author explained the fundamental concept of the hedonic method. The main contribution of the paper

was to introduce the time dummy variable for the hedonic model of cloud price to analyze AWS' cross-sectional data between 2009 and 2015 (See Figure 2—22).

$$h_{DV}(x_i)\colon \ln P_{i,t} = \alpha + \sum_k \beta_k X_{k,i,t} + \sum_t \delta_t D_{i,t} + \varepsilon_{i,t}$$

$$h(x_i) = E[p_i|x_i] = E[mc(\cdot)|x_i] + E\left(\frac{\boldsymbol{D_i}(\cdot)}{\partial \boldsymbol{D_i}(\cdot)/\partial p}|x_i\right), \qquad D_i(\cdot) = D\left(x_i, p_i, x_{-i}, p_{-i}; A\right)$$

$\boldsymbol{h_{DV}(x_i)}$: Hedonic values with a dummy variable
$\boldsymbol{P_{i,t}}$: Price of goods "i" at time "t."
$\boldsymbol{mc}(\cdot)$: marginal cost
$\boldsymbol{D_i}(\cdot)$: The demand function for good i
x_i Hedonic Characteristics of cloud services
DV: Dummy variable
α: Constant value
β_k: Coefficient of characteristics
(x_{-i}, p_{-i}): Characteristics and price of other goods
δ_t: Time dummy coefficients
$D_{i,t}$: Time dummy variable
$\varepsilon_{i,t}$: Error term of the regression equation
k: Total number of Characteristics
t: Number of the time period (year)
$X_{k,i,t}$: The vector of cloud characteristics
A: Consumer preference over characteristics

Figure 2—22 Hedonic Function with Time Dummy Variable

According to their experiment results, the adjusted R2 value was 0.9792 for 277 data points. In comparison with other papers, their work made a significant improvement. However, the author could not collect enough data points for earlier years of AWS cloud service. It might explain that the author did not provide the coefficient results for time dummy variables. The calculated result for the time dummy effect has a big gap. Furthermore, the *p*-value of storage is less significant than other cloud service characteristics. The value of the storage coefficient showed as negative. As the author concluded, the major issues of the paper are 1) a small sample of data is not enough to lead a reasonable conclusion, and 2) some hidden cloud characteristics were left out.

All the above issues regarding of hedonic pricing model have been solved in Chapter 3. This study developed a much-sophisticated hedonic pricing model for cloud services. The model categorized hedonic values with three types of cloud characteristics or three variables, namely intrinsic, extrinsic and time dummy (See Figure 2—23). It improved the accuracy of the future cloud price. The significant contribution of the study is to unveil a depreciation rate of cloud service, which is equal to 20%. This rate is equivalent to Moore's law for computer hardware.

In addition to the hedonic method, there are also many other value-based pricing models. For example, Jain's [88] social welfare pricing model focuses on the sum of cloud consumers' value. The performance-based pricing model [89] is associated with cloud resource and application risks. Feature-based pricing [90] that is related to prioritizing cloud features. The service-based pricing model [91] correlates to the Service Level Agreement (SLA).

Extrinsic Value

$$ln[p(X)] = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=1}^{l} \xi_j z_j + \sum_{t=1}^{T} \delta_t d_t$$

Intrinsic Value Time Dummy

p: price of cloud service
X: the vector of cloud Characteristics
β_0: interception coefficient value
β_i: Coefficient of cloud intrinsic characteristics
x_i: Hedonic intrinsic characteristics
i: the number of intrinsic characteristics
k: the total number of intrinsic characteristics
ξ_i: Coefficient of cloud extrinsic characteristics
z_j: Hedonic extrinsic characteristics
j: the number of extrinsic characteristics
l: the total number of extrinsic characteristics
T: the total length of time period
d_t: the variable of time dummy
δ_t: Coefficient of time dummy
t: number of the time period (year)

Figure 2—23 A Comprehensive Cloud Pricing Model by Hedonic Analysis

Jain's model is much similar to a spot pricing model. In other words, cloud users can submit their ceiling bid prices (willingness to pay) and CSP can adopt different algorithms to schedule and allocate cloud resources based on the optimized metrics (such as profits, cloud capacity, performance, time of a day, energy consumption, etc.). However, it is quite challenging to be implemented because it left out the cost components of the cloud services. Naturally, all customers would like to have free or near-free cloud resources, but "cloud computing will never be free" [92].

Lucanin's [89] performance-based price is mainly driven by CPU's properties, namely electricity price, and CPU's temperature traces. The paper claimed that it could save up to 32% of the cost under certain assumptions. The pricing model is that the cloud price is dependent on the workload characteristics and determined by the performance that is perceived by users. However, the cloud price is not only dependent on the CPU but also memory, storage size, access bandwidth, and other service characteristics.

Kar's [90] feature prioritized pricing model is to estimate the potential value of the workload to the individual user for a particular context. The paper proposed an integrated approach to price IaaS resources from a multi-user perspective. In other words, the model will aggregate all potential values for all cloud features. The issue is how to define the benefits of these cloud features from a cloud customer perspective because these values are highly subjective.

Wu et al.'s [91] SLA-based model is a resource allocation or scheduling for SaaS delivery. Similar to the feature-based concept, SLA can be interpreted as different cloud features, which include response time, provisioning time, data transferring speed, etc. However, SLA is not only response time and data transmission speed, but also include security, cloud regions, and zone diversity, API compatibility, auto-scaling, vertical and horizontal scaling without a reboot, burstable CPU, backup-snap, 24X7, etc. Many of these features are quite challenging to be measured by the cost-based pricing. They are built into cloud service as a whole for a particular CPS to differentiate its service from other CPS competitors.

Despite many theoretical models that are illustrated above, AWS first launched the new pricing model in 2014, namely the Lambda function. It is delivered by the serverless sandbox technology, which is also known as Function as a Service (FaaS). It is supported by the Docker container and Application Programming Interface (API). A Docker is the default container runtime engine, and a container can be easily destroyed, stopped and built with minimum effort of set-up and configuration or "ephemeral," which is like a sandbox.

Adam Eivy [93] argued that the serverless sandbox allows cloud consumers to have infinite cloud resources with vendor-free. In other words, if all CSPs support Open API, cloud users can quickly switch among the different CSPs without worrying about vendor-locked in. The price of AWS Lambda function consists of two components, namely, Hit Pricing and Compute Pricing (Memory allocation). AWS [94] and Peter Sbarski showed [95] the details of how to calculate the total cost of AWS Lambda function. We can use Equation 8 to calculate the AWS Lambda price. (Equation 2-5)

$$P_t = h_r + m_r = (\alpha\lceil X_{100}\rceil h - k) \times r_h + \left(\frac{\alpha\lceil X_{100}\rceil h}{10}\frac{R}{y} - g\right) \times r_m \qquad (2\text{-}5)$$

where, P_t is the total price of the Lambda function per month, $h_r$is the hit price and $m_r$is a memory resource price. a is the constant value of second per month = 2,628,000. $\lceil X_{100}\rceil$ is a

ceiling function for the round-up integer of code execution time/per 100ms. h is the hit rate/per second (execution rate of computing request). If the user's code execution time is less than 100 ms, for example $X_{100} = 85ms$, "$\lceil X_{100} \rceil$" is equal to or normalized to 1, if the code's execution time is more than 100, e.g., X_{100} =101, "$\lceil X_{100} \rceil$ is equal to 2. R is the allocated memory resource, e.g., 256 MB-second. y is the baseline memory 1024MB-second (reference price). $r_h$is the price rate \$.020/per million hits (Lambda@Edge. r_h= \$0.6). r_m is the price rate =\$1.667E-06/per 100ms for 1024Mb-s. k is the free allowance of the first one million hits/per month (If the hit rate is less than about 23 hits/per minute, it would be free for compute resource. However, Lambda@Edge has no free allowance). g is the free allowance of 1024Mb-s is 400,000 GB-second/per month. For instance, if a cloud user has an application code that has 50 hit/per second and code execution time is 125 ms, and the memory size is allocated to 256MB/per 100ms, we should have h=50, $\lceil X_{100} \rceil$=2, y =1024MB/per 100ms, The total monthly bill is:$P_t = h_r + m_r = (\alpha \lceil X_{100} \rceil h - k) \times r_h + \left(\frac{\alpha \lceil X_{100} \rceil h}{10} \frac{R}{y} - g\right) \times r_m = (2{,}628{,}000 \times \lceil 125_{100} \rceil \times 50 - 1{,}000{,}000) \times 0.0000002 + \left(\frac{2{,}628{,}000 \times \lceil 125_{100} \rceil \times 50}{10} \frac{256}{1024} - 400{,}000\right) \times 0.000001667 = \$52.36 + \$10.95 = \$63.31/\text{per month}.$

However, if we can reduce the execution time of code to less than 99ms, the monthly bill can drop down \$31.56/per month. From a CSP perspective, this pricing model allows CSP to allocate 75ms (200ms – 125ms) to compute execution time for another user. On the other hand, the cloud consumers only pay what the code execution time slot, which is Pay As You Use (PAYU) or Pay per Task (P/T). Obviously, this price does not include the cost of storage, API gateway, and data egress.

The trend of the cloud pricing model is moving towards a much more flexible direction, and the billing method becomes PAYU and P/T rather than upfront lump sum payment. However, the bad news for this model is if the number of hits/per second is remarkably higher, the cost of the Code of Demand (CoD) could be out of a hand. Sometimes, it could be three times higher than VM on-demand [93]. Overall, the new pricing model is to support the new service (FaaS) that is working with a new platform or orchestration, such as AWS' Cloud-Watch Rackspace's OpenStack, and Google's Kubernets. Following AWS' step, both Google Cloud Platform (GCP) and Microsoft Azure also launched Functions as a Service (FaaS) platform in early 2016. All three CSPs have almost the identical pricing model for serverless computing (See Table 2—8).

Table 2—8 FaaS Pricing Model

CSP	Free Tier (per month)	Memory Resource Allocation *	Price/per 128MB/per 100 ms
AWS	400,000 GB-s	1024 MB	$0.000 0166 7/GB-s
	1 million executions		$0.20 per million executions
Google Cloud Platform	400,000 GB-s	1024 MB (1.4GHz CPU)	0.000 0165 0/GB-s
	2 million executions		$0.40 per million executions
Microsoft Azure	400,000 GB-s	Up to 1,536 MB	$0.000 0160 0/GB-s
	1 million executions		$0.20 per million executions

*Note: Different sizes of memory allocation have different prices/per 100 ms execution. Here, we only use 1GB memory as an example.

2.4.5. Summary

This chapter reviewed the number of papers regarding cloud pricing models from 2008 to the present. It presented an in-depth analysis of these research works, which can be summarized from three basic pricing strategies according to the principle of value theory. We highlight these pricing models in Table 2—9.

Table 2—9 Summary of Cloud Pricing Models Survey

Category of pricing models	Mathematical Equation of model	Main Contributions	Potential and Gaps
Marketed Based pricing: Toosi et al.'s Max CSP Revenue (2014)	$\pi = \max_{r_t} \sum_{t=0}^{\Gamma-1} r_t\varphi + \alpha p u_t(l_t^r + r_t) + p(l_t^o + o_t) + \beta p(l_t^s + s_t)$	Novelty Idea of how to maximize the cloud revenue for the fixed cloud capacity. It combines all revenue streams including on-demand, reserved and spot instance	Omitting the cloud cost could be an issue in practice. It is challenging to define a unified price practically. GCP and AWS have different charging mechanisms. AWS can empty spot instances at any time and only gives two minutes advance warning time.
Marketed Based pricing: Xu et al.'s dynamic pricing model (2013)	$J^*(x,t) = \sup_{u\in\mathcal{U}}\left(E_u\left[\int_0^t p(s)dX(s)\right]\right), \forall t > 0$ $\frac{\partial J^*(x,t)}{\partial t} = \sup_p[px + f(p)J^*(x+1,t) - J^*(x,t) - g(p)J^*(\Delta x,t)]$	The main contributions of this paper offer an alternative pricing model for CSP to price its spot instance dynamically	However, the spot pricing cannot be generalized to all instances. In one case, the spot price reached a ridiculously high price - $999.00. Usually, the spot instance price variation with time is neither convex nor continuous. Two critical functions are defined as more like a power function rather than a Poisson distribution function
Marketed Based pricing: Orna, et al. Traceable data (2013)	$\delta_i = -a_1\delta_{i-1} + \varepsilon(\sigma),\ and\ p_i = p_{i-1} + \delta_i$	It intended to unveil the spot price mechanism of AWS and indicated spot price within a limited band	If the authors adopt the auto-regression or statistical method, the result and conclusion may have more weight when the p-value is demonstrated.

Marketed Based Pricing: Zheng, Liang, et al. (2015)	$\max_{i=1,\cdots,M} T_i F_{p(t)}(p_b) = \frac{t_s + t_o - Mt_r}{M\left(1 - \frac{t_r}{t_k}\left(1 - F_{p(t)}(p_b)\right)\right)}$	1.) Price orientation bid strategy, 2.) SLA priority bid strategy, and 3.) MapReduce workload application. 4.) Based on the authors' observation, they conjecture that only a few users bid for spot instances due to heavy-tailed spot price distribution	1) In practice, the bid price could exceed the on-demand price. 2.) The maximum revenue function analysis did not include the cost from a CSP perspective. 3.) The assumption of uniform distribution for bid prices is a contradiction with the later contents 4.) It is an unrealistic assumption to isolate the spot price alone.
Cost-Based Pricing Greenberg et al. (2008)	Not Applicable	It highlighted a significant issue across many data centers at that time (before 2008). It provided a rough estimation of cloud data center element cost	It ignored space costs, which could be up to 15% of the total cost. The assumption of electricity power cost was at the lower end.
Cost-Based Pricing Walker, Edward (2009)	$R_p(purchase) = \frac{NPV}{NPC} = \frac{\left(1 - \frac{1}{\sqrt{2}}\right) \times \sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T}}{\left(1 - \left(\frac{1}{\sqrt{2}}\right)^Y\right) \times TC}$ $R_l(lease) = \frac{\sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T}}{Y \times TC}$	The paper highlighted a significant issue across many data centers at that time (before 2008). The authors identified some approaches to improve data center efficiency.	The primary assumption of the future CPU price is stable, but the real future CPU price in the market is very volatile. Subsequently, the expected NPV value is a probability distribution among a specific range
Cost-Based Pricing Walker, Edward (2010)	$\Delta NPV = \sum_{T=0}^{N} \frac{C_T - E_T + L_T}{(1+I_F)^T} + \frac{S}{(1+I_F)^N} - C$	The main contribution is to demonstrate how to use the NPV concept to construct a cloud cost-based model by taking consideration of Moore's law. The author provided a particular period for the decision of on-premises or off-promises	The predicted cost per Gbytes is dependent on previous observation. Different sources of price data collection could lead to different NPV results. As a result, the range of ΔNPV value could be uncertain.
Cost-Based Pricing, Xu, Hong, and Baochun Li (2013)	$\max_{p*} \int_{v_0}^{v_1} R_v(p) f(v) dv,$ $s.t. \int_{v_0}^{v_1} D_v(p) f(v) dv \le C$ $S_v(p) = vU[D_v(p)] - pD_v(p) = \frac{\alpha p}{1-\alpha}\left(\frac{v}{p}\right)^{\frac{1}{\alpha}}$ $S_v(p) \ge 0, \forall v, \ Over\ p$	The major contribution of their work was to introduce a probability density function $f(v)$ for cloud market demand.	1.) The revenue optimization without a cost component appears to be not obeyed to the basic economic principle. 2.) The price model remains as a theoretical discussion 3.) the assumption of elasticity value that is equal to 0.3 requires further explanation.
Cost-Based Pricing, Joe-Wong et al. (2012)	$CS_j = \max_{x_j^*}\left[U_j(x_j) - r_j x_j^\gamma\right]$ $\pi_b = p\sum_{j=1}^{n}\left(\mu_j x_j^*\left[\mu_j^\gamma p\right]\right)^\gamma$	The significant contribution of the paper is that it adopted the iso-elasticity function for the utility to model the cloud resource price. It emphasized that CSP could leverage cloud customers' surplus to maximize its revenue.	1) Various bundle types for cloud resources seem only to add complexity to cloud pricing models. 2) The model did not clearly articulate two different meanings of economic utility and the capacity utility (or utilization rate) for fairness. 3) The assumption for all customers' utility functions that are continuous and concave may need further consolidation.
Cost-Based Pricing, Shahrad et al. (2017)	$\boldsymbol{E(p) = E(R) - E(Y)}$	It is the first time to propose a cloud price model with a self-capping solution to help CSP to increase the utilization rate of cloud infrastructure capacity	The profit functions of cloud customers have to be differentiable. Otherwise, the optimal capacity value cannot be found.
Value-Based Pricing	$BA_p = \beta_{0p} + \sum_{i \in I} \beta_{ip} x_i + \varepsilon_p, [p \in P]$	It is the first time to apply the hedonic method for cloud computing prices.	The interpretation of their experiment results seems to be inaccurate. The adjusted R square value is only between

El Kihal, Siham et al (2012)			43%~69%. It means the linear regression is not "goodness of fit."
Value-Based Pricing, Zhang, Liang (2016)	$h_{DV}(x_i)$: $\ln P_{i,t} = \alpha + \sum_k \beta_k X_{k,i,t} + \sum_t \delta_t D_{i,t} + \varepsilon_{i,t}$	The issue of "goodness of fit" was picked by Zhang's hedonic regression formula (semi-log form). The other significant contribution is to introduce a time dummy variable in the hedonic analysis for cloud price.	1.) The hedonic method that some hidden cloud characteristics were left out. 2.) The coefficient of time dummy variables between 2009 and 2015 was not provided.
Value-Based Pricing, Wu. et al. (2018)	$ln[p(X)] = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=1}^{l} \xi_j z_j + \sum_{t=1}^{T} \delta_t d_t$	-20% of Cloud services depreciation rate that is equivalent to Moore's law for computer hardware. The prediction for future cloud price has been significantly improved	Some extrinsic coefficient values require further consolidation when all leading CSPs data become available.
Value-Based Pricing Adam Eivy, and Peter Sbarski (AWS Lambda Function) (2017)	$c_t = \square_r + m_r = (anx - k) \times r_{\square} + (bnxy - g) \times r_m$	Cloud users pay precisely the code execution time or Code on Demand if the code the execution time is very close but less than the unit 100ms. CSP can allocate unused execution time for other cloud users	If the code execution time is unknown or very long, the monthly bill can be quickly out of hand. Sometimes, the price could be three times more than VM on-demand pricing model

Based on the above table, we notice that the goal of cloud pricing models is to maximize business profits and to improve cloud resources efficiency, which is to minimize cloud infrastructure costs. The common trait of early solutions was the cloud infrastructure cost is the central theme of the pricing model. During the early era of cloud computing, many researchers mainly focused on the cost and limited capacity of cloud infrastructure. Walker's two papers [77] [78], Greenberg's [60] and Joe-Wong's [113] [114] studies provided a good example. When the cloud market becomes the mainstream computing resource, especially after AWS launched the spot-instance in 2009, the research topic had been shifted to market-based pricing. Xu [80] [56], Orna [61], and Toosi [55] included on-demand, reserved and spot-instance models into their solution of profit maximization. Just recently, El Kihal [84], Mitropoulou [85], Zhang [86], and Chapter 3 adopted the hedonic method to evaluate CSP's pricing, which is considered as a value-based pricing strategy.

The differences of three pricing strategies are that value-based pricing is driven from the demand side, while cost-based pricing is oriented by the supply side and the market-based pricing is to focus on the equilibrium of supply and demand. The primary goal of having different pricing strategies and generating multiple price models is to capture more surplus value under a cloud customer's demand curve. From Figure 2—24, we can see that moving from 2009 (diagram A)

to present (diagram B), more customer surplus values have been gained after new cloud service features alongside new price models have been created. For example, the dedicated host pricing models supported by new cloud technology, such as Cloudheat [109] can draw customer's surplus values from the upper end (area "1") of the demand curve, and discount pricing category enabled by desktop grid technology can draw a customer's surplus value from the lower end (area "2"). It implies that the new cloud price models underpinned by innovative cloud technologies can maximize cloud business profit for CSP.

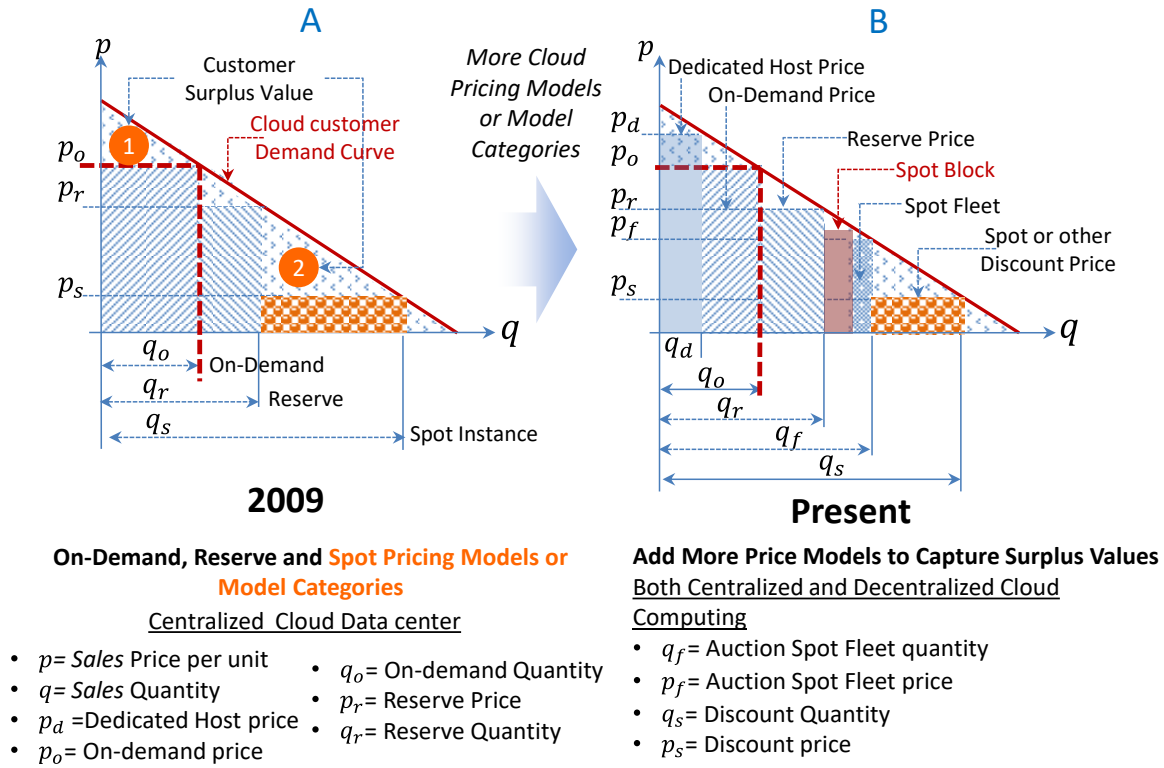


Figure 2—24 More Pricing Models to Capture Customer's Surplus Value

Chapter 3

Hedonic Pricing of Cloud Computing Services

Cloud service providers (CSP) and cloud consumers often demand to forecast the cloud price in order to optimize their business strategy. However, the pricing of cloud services is a challenging task due to its services complexity and dynamic nature of the ever-changing environment. Moreover, the cloud pricing based on consumers' willingness to pay (W2P) becomes even more challenging due to the subjectiveness of consumers' experiences and implicit values of some non-marketable prices, such as a burstable CPU, dedicated server, and cloud data center global footprints. Unfortunately, many existing pricing models often cannot support value-based pricing. This chapter proposes a novel solution based on value-based pricing, which does not only consider how much does the service cost (or intrinsic values) to a CSP but also how much customer is willing to pay (or extrinsic values) for the service. This study demonstrates that the cloud extrinsic values would not only become one of the competitive advantages for CSPs to lead the cloud market but also increase the profit margin. The approach is referred to as a hedonic pricing model. This chapter shows that the hedonic model can capture the value of the non-marketable price. This value is about 43.4% on average above the baseline, which is often ignored by many traditional cloud pricing models. This work also shows that the Average Annual Growth Rate (AAGR) of Amazon Web Services' (AWS) is about -20.0% per annum between 2008 and 2017, ceteris paribus. In comparison with Moore's law (-50% per annum), it is at a far slower pace. This chapter argues this value is Moore's law equivalent in the cloud. The primary goal of this research is to provide a less biased pricing model for cloud decision-makers to develop its optimizing investment strategy.

3.1 Introduction

PRICING cloud computing has always been a big challenge not only to many Cloud Service Providers (CSPs) but also to many cloud consumers because of the exponential growth of new service features or characteristics that appear almost daily. Although pricing of cloud service delivery has often been drawn an analogy as a new public utility service [125], the underlying

---

This chapter is derived from

- **Caesar Wu**, Adel N. Toosi, Rajkumar Buyya, and Kotagiri Ramamohanarao, Hedonic Pricing of Cloud Computing Services, IEEE Transactions on Cloud Computing (TCC), ISSN: 2168-7161, IEEE Computer Society Press, USA (in press, accepted on July 15, 2018)

structure of cloud pricing is much more complicated than the traditional public utility services due to the rapid development of cloud technologies and multiple layers of service delivery models (or Anything as a Service: XaaS).

As Weinman [1] had noticed, the utility pricing or Pay-As-Your-Go (PAYG) is not the only possible model for the cloud. Some firms have begun to explore their marketing strategy to support "pay-what-you-like." He indicated one of the important lessons that CSPs should learn from other industries is that relying on innovative cloud services and technologies is not enough. CSP has to also come up with new pricing models for its services. This means that CSPs should "move beyond competition just on a price to compete on pricing." The question of how to move beyond competition just on price leads to the idea of how to establish innovative pricing models for cloud services. The primary objective of the cloud pricing model is to capture cloud service values along with its pricing variation as well as the dynamic nature of cloud technology development.

Our observation shows that the revenue growth of Amazon Web Services (AWS), one of the leading global CSPs, has a positive correlation with its cloud characteristics (See Figure 3—1). This means various cloud service features, such as PAYG, burstable CPU, data center global footprint, GPU, one account for all location, etc. (Notice that the number of characteristics has been increased from just a few in 2006 to more than a thousand in 2017 due to AWS' continuous cloud innovation [145]). The fundamental question is, "Will the cloud characteristics impact its service price or customer willingness to pay (W2P)?" If so, what is the relationship between cloud characteristics and service prices? Most importantly, how we can calculate or estimate the values of these characteristics. One of the solutions is a so-called hedonic model. The compelling reason to propose the hedonic model is that it can capture non-market values (extrinsic values) for the cloud ecosystem and evolutionary characteristics that either directly or indirectly impact on its service prices.

Empirically, the basic premise or assumption of the hedonic function is that the product price difference is closely aligned with its characteristics (or features) variation. This means that if we can successfully establish a relationship between cloud service price differences with various cloud service characteristics, we will be able to estimate the price of cloud services accurately.

Another advantage to consider the hedonic approach is that the cloud price can be modeled by the regression analysis for the cloud service features along with its price variation over a period. In comparison with other methods, such as survey-based or contingent valuation [126] or Delphi [127] method, the hedonic regression approach is quick and cost-effective if the chosen dataset is sufficiently large for the regression analysis. Moreover, it can be easily updated. It is a good fit for the cloud environment because of its ever-changing market conditions and rapid technological innovations.

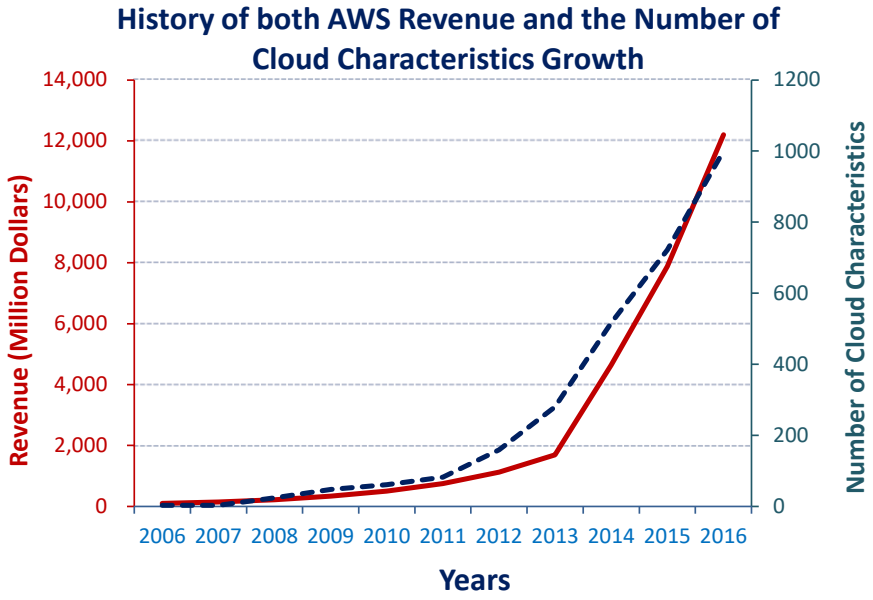


Figure 3—1 AWS Revenue Expansion and Characteristics [145]

Historically, the hedonic model has two different objectives. One is to predict the future price of goods or services that customers are willing to pay. The purpose of hedonic prediction is to help decision-makers to make an optimized strategic decision. The other is a hedonic index, which is to establish a price ratio by comparing it with a price in a base period. The goal of the hedonic index is to monitor the price of either inflationary or deflationary, which is to verify what has happened in the past.

In this chapter, we mainly focus on hedonic prediction or estimation. In order to achieve a better estimation, we introduce the concept of both intrinsic and extrinsic variables for the hedonic function model, which is inspired by G.E. Moore [20] as a solution for the cloud pricing problem. The intrinsic variables of cloud instances are defined as cloud resources, such as memory, CPU, storage, and network performance. They often appear as numerical variables. In contrast, the

extrinsic variables can be anything from Burstable CPU, OpenStack compatible API, the global footprint of Cloud Data Center (DC), Mobile Application, vertical scaling without a reboot, to even one account for all locations. They are binary or categorical variables. In this chapter, we propose a pricing model based on hedonic principles to capture the values of both intrinsic and extrinsic variables. This can help both CSPs and cloud consumers to estimate cloud prices more accurately. In addition, it explains the reasons why some market leaders of CSP do not only compete based on the price of intrinsic value but also on the price of an extrinsic one. Our proposed model will help many cloud decision-makers to understand price differentiation. We believe it will become a practical tool in a price modeling toolbox for many CSPs and it will also provide a pricing technique for many cloud consumers to select the right CSP for their application need. In summary, we have made the following contributions:

1. We articulate that cloud prices are dependent on both intrinsic and extrinsic variables according to the utility theory. We have also demonstrated how to compute these extrinsic values practically.
2. We construct a novel form of hedonic function for cloud pricing, which consists of three explanatory variables: intrinsic, extrinsic and time dummy.
3. To the best of our knowledge, this is the first attempt to use the time dummy variable to correctly calculate the Average Annual Growth Rate (AAGR) for cloud service. If we use AWS as a benchmark, it is about -20.0% per annum. This rate basically captures Moore's law behaviors. It is also the first time to comprehensively describe the context regarding the hedonic model for cloud pricing. Moreover, it attempts to pricing cloud services with both panel and cross-sectional datasets.
4. We show that cloud price is declining but at a slower pace than what Moore's Law predicts for computing hardware [158]. We argue this slow pace is due to the non-marketable pricing values (by alone, these features have no value), namely, extrinsic variables or characteristics.
5. We exhibit that our novelty pricing model can provide a good and simple solution to predict cloud prices. We also show that a customer is paying more than a typical baseline service price (a standard configuration of cloud instance) on average for their business needs.

This study uses AWS data in 2014 to generate a simple hedonic regression model. Based on this model, we estimate a cloud price (by a typical configuration of cloud instance) in 2017 and

then compare it with the real price in 2017. Our results show that the model can predict with an average accuracy of 87%. We use AWS 10 years unbalanced panel (longitudinal) data to construct a hedonic model with time dummy variables. According to this model, we can calculate the value of AAGR. By using AAGR, we can revise our estimation of cloud price. However, this price estimation does not take into consideration of the extrinsic variables. In order to capture the extrinsic values, we develop a comprehensive hedonic model to calculate the value of each extrinsic characteristic based on the cross-sectional data of five CSPs. Finally, we update the estimated cloud price to achieve many accurate results based on a particular type of workload.

The rest of the chapter is organized as follows: Section 3.2 provides the background information. Section 3.3 reviews related works and introduce the hedonic concept. It consists of three parts: the empirical work of hedonic analysis, the hedonic pricing model for computer prices and the hedonic model for the cloud. Section 3.4 defines the hedonic function for cloud pricing. Section 3.5 provides a performance evaluation. Section 3.6 analyses the results with detailed discussion. The final section provides summary information.

3.2 Background

To set the background, we consider a scenario where a Chief Information Officer (CIO) of a firm needs to make a strategic investment decision whether to build their own private cloud (on-premises) or just migrate IT workloads to the cloud provider (off-premises, either private or public cloud infrastructure). Assume that the firm has its own on-premises IT infrastructure that still supports its existing business applications and the book value of IT assets that cannot be written off for the next 12 ~ 36 months.

In this discussion, we ignore other issues such as types of IT workload, migration cost, and system lifecycle management (SLCM) cost. The fundamental issue can then be boiled down to "how can we estimate the future market price of cloud services for the next 12~36 months?". The logic behind this line of reasoning is if we can successfully predict or estimate the cloud price along with its service features (or cloud characteristics) that the business requires, we can select either building or buying or a hybrid solution for IT infrastructure. This means that if we can use the pricing model to predict the future price of cloud services accurately, we can help the CIO to

develop a better IT investment strategy. However, cloud pricing modeling is much more complicated due to the hedonic nature of many of its characteristics or features.

The term "hedonic" or "hedonism" was derived from a Cyrenaic parable in ancient Greek. It literally means "The Choice of Pleasure" [128] in contrast to "pain." Economically, the connotation of hedonic is the meaning of gain, which is the opposite of losing. From a cloud consumer perspective, hedonic values can be interpreted as some implicit benefits that are derived from specific cloud characteristics offered by a particular cloud service. Often, these service values are not only dependent on its intrinsic variables but also on many extrinsic variables.

Traditionally, the price of any given cloud service (typically IaaS) is often determined by its cost components or required resources. It is referred to as cost-based pricing. With cost-based pricing, one of the disadvantages is that it cannot capture many cloud service characteristics. The other conventional approach to pricing is based on supply and demand, which is dependent on the market competition or the existing market conditions. We often call it market-based pricing. Unfortunately, many innovative services and cutting-edge technologies do not have an existing market to decide the price of goods. In contrast, the hedonic pricing model can overcome these issues to some extent because it can capture both intrinsic values (resource costs) and extrinsic values (service characteristics) and can estimate the missing or future price based on the existing market [152][163]. This way, we can present a hedonic based pricing model to CIOs to estimate the future cloud price accurately. (Table 3—1 lists all the acronyms used in this chapter.)

Table 3—1 Acronym Used in This Chapter

Acronym	Definition	Acronym	Definition
AAGR	Average Annual Growth Rate	I/O	Input / Output
API	Application Programming Interface	IaaS	Infrastructure as a Service
AWS	Amazon Web Services	OLS	Ordinary Least Square
CAGR	Compound Average Growth Rate	PAYG	Pay As You Go
CIO	Chief Information Office	RAM	Random Access Memory
CSP	Cloud Service Providers	SLCM	System Lifecycle Management
EBS	Enterprise block Store	SSD	Solid State Drive
EC2	Elastic Compute Cloud	vCPU	Virtual Central Processing Unit
ECU	Elastic Compute Unit	VM	Virtual Machine
GCP	Google Cloud Platform	W2P	Willingness to Pay
GPU	Graphics Processing Unit	XaaS	Anything as a Service
HDD	Hard Disk Drive	YoY	Year on Year

3.3 Related work

The modern hedonic theory can be traced back to the founder of modern utilitarianism, Jeremy Bentham [129]. In Bentham's view, the hedonic value is a sensational pleasure. He identifies seven main variables (IDCNPFE) to calculate hedonic values. We show these values and their relevance to cloud computing values in Table 3—2

Table 3—2 Bentham's Seven Hedonic Variables Relevant to Cloud

Bentham's Hedonic Variables (IDCNPFE)	Bentham's Definition	Value Range	Hedonic Values Relevant to Cloud
Intensity (I)	the amount of quality for pleasure or pain	0-10	Quality of Services
Duration (D)	how long the pleasure or pain will last	From minutes to weeks	Usage Time
Certainty (C)	the probability of the pleasure or pain will occur	0 – 100%	The certainty of a price discount
Nearness (N)	how far off in the future	Now – Years	When discount price starts & ends
Purity (P)	how the decency of pleasure	0-100%	Dependent conditions to obtain cloud service
Fecundity (F)	the probability of reproducing the pleasure or other pleasures	0-100%	The probability of having discount price & more cloud service features continuously in future
Extent (E)	the number of people will be impacted by the pleasure	One or Many	Number of people can share the Cloud services

In contrast to Bentham's view, John Stuart Mill [130] emphasized a higher level of intellectual happiness, which differs from Bentham's pure hedonic value. He stated, "It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied." In today's cloud pricing term, Mill's hedonic value means to pursue a good result for business applications, while Bentham's hedonic value emphasizes to maximize the number of cloud service characteristics for the maximizing number of cloud customers.

G.E. Moore [20] divided the hedonic values into two kinds: intrinsic (or non-instrumental) and extrinsic (or instrumental) [131] . This is Moore's significant innovation to hedonic theory. The concept of intrinsic value means that something is good or valuable on its own and the value is independent of others. For example, RAM, CPU, and storage can be considered as intrinsic values. In contrast, the extrinsic value is determined by the relationship with others, such as PAYG,

burstable CPU and 24X7 supports, which are dependent on RAM and CPU. G.E. Moore's concept of intrinsic and extrinsic values underpins our hedonic model.

3.3.1 The Empirical Hedonic Analysis

The empirical hedonic analysis had been adopted as early as the 1920s. Zvi Griliches [132] generalized the hedonic regression model along with a semi-logarithmic form for the vehicles' application in the 1960s. Griliches noticed many practical issues of the hedonic model analysis [151]. One of them was, "How should the regression framework be expanded, what variables should be added to it, so as to keep the resulting estimates stable in facing of changing circumstances?". He emphasized the essence of hedonic analysis is to estimate the "missing" prices or values due to quality or characteristics change, which influences our hedonic models for the cloud pricing.

3.3.2 Hedonic Model for Computer Price

In addition to the property and automobile applications, another popular application of the hedonic model is computer hardware, such as a mainframe, workstation, and personal computer. Since the later 1970s, there have been countless publications regarding of hedonic price index of workstation and Personal Computer (PC). One of the earlier works was contributed by R. Michaels' [133]. He demonstrated how to establish a hedonic function with CPU performance, memory size, the speed of I/O, storage capacity and high-speed storage characteristics plus brand name and time dummy variables. Based on the regression analysis, the paper indicated that brand name had an insufficient impact on implicit prices and the deviation of quality-adjusted prices is smaller for the high-end computer equipment. The main conclusion of the paper was "observed price variations to be consistent with the economic theory" (value for money).

For the same topic, Cole et al. [134] presented and compared different PC hedonic price indexes with a matched-model index and demonstrated that the traditional matched-model index is inadequate for PC products because the index excluded many new replacement PC models due to rapid technology improvement in the PC industry. However, the authors did not give an explanation for why was the reason for PC price deflation.

Ernst R Berndt and Zvi Griliches [135] separated the price-decreasing problem into two issues: one is a price index and the other is the ratio of performance against price. They provided a variety

of price indexes to serve the purpose of the deflation explanation for the microcomputer. The indexes were a kind of benchmark to measure "a technological frontier in the PC market" based on an unbalanced panel data. The paper reported testing results with various hedonic regression models, especially leveraging many dummy variables, such as year, vintage, process bit-length and age of PC. One of the apparent results was the PC price was decreasing, although the quality of the PC was improving. Moreover, the authors noticed the issue of the parameters of the regression model has high variances and is unstable. The decision to select a set of variables from a pool of characteristics was arbitrary.

In contrast to many indexes oriented hedonic analysis, Rao et al. [136] mainly addressed the issue of how to economically analyze information systems (IS), which is how to acquire workstation hardware in the 1990s for many large organizations. The authors presented a hedonic function in the Box-Cox [137] transformation form (Equation 3-1) in order to extract a pattern between prices and the hardware characteristics.

$$\frac{y^{\lambda}-1}{\lambda} = \beta_0 + \beta_1 \frac{x_1^{\lambda}-1}{\lambda} + \beta_2 \frac{x_2^{\lambda}-1}{\lambda} + \cdots + \beta_n \frac{x_n^{\lambda}-1}{\lambda} \tag{3-1}$$

where y is the workstation price, x_i^{λ} is the workstation ith characteristic. $\beta_1 \cdots \beta_n$ are the coefficients, β_0 is the intercept value. λ is the transformation power parameter. The authors had noticed there were many difficulties in constructing a hedonic function form, some of which still exist for determining cloud service pricing. These issues include:

1. How to aggregate the characteristics of a good or service at a box level.
2. How to specify the characteristics in detail,
3. How to select each characteristic that can reflect both customers' values and resource costs,
4. How to handle the evolutionary characteristics,
5. How to trace and measure these characteristics at the box level,
6. How to apply the hedonic model or appropriate hedonic function at the box level,

In comparison with Rao's hedonic model, Pakes' paper [138] demonstrated a relatively easy way to construct a hedonic model from an index perspective. Parkes' empirical results show that PC's hedonic price had a sharp decline while the traditional matched model exhibited the near-zero values. According to Hulten [139], Pakes made three significant contributions to the hedonic analysis:

1. The coefficients of the hedonic function are not always fixed over time. Moreover, the sign of the coefficient is not necessary to be positive. In other words, some product's characteristics may have a negative impact on the overall hedonic values.
2. Two hedonic functions of the same product could be different from each other.
3. Each hedonic function is sufficient to make a quality judgment.

In addition, Pakes' theory of hedonic function is much easier to be grasped in comparison to other forms that have too many "restrictive assumptions." It can be directly derived from the theory of microeconomics [140], in which the hedonic price reflects the price elasticity. Let (x_i, p_i) denote the characteristics and the price of the product "i" and Q_i is the quantity of demand of the product. Note Q_i is dependent on the price p_i and x_i. We can graphically show the product's price in Figure 3—2

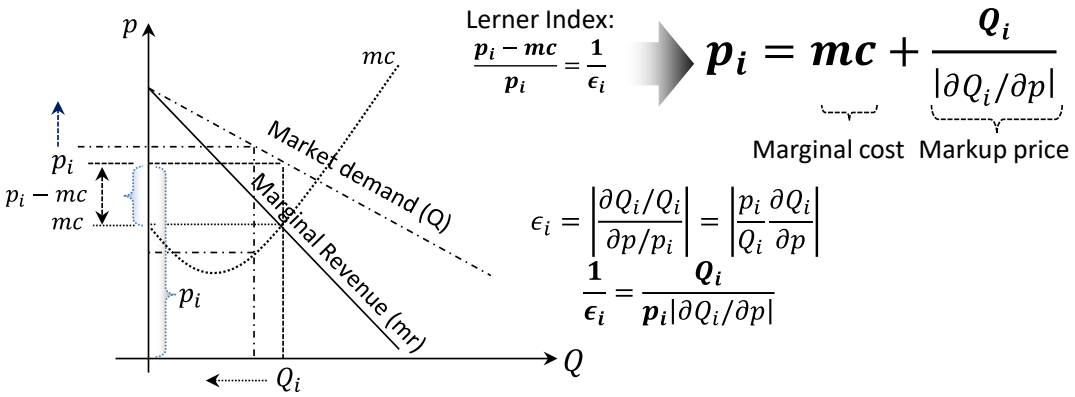


Figure 3—2 Theoretical Interpretation of Hedonic Price

From the Lerner index, we should have the Equation 3-2 [7]

$$L_i = \frac{p_i - mc}{p_i} = \frac{1}{\epsilon_i} \tag{3-2}$$

where "p_i" is a market price of the product, and "mc" is the marginal cost for the product and ϵ_i is the elasticity. From microeconomics theory [140] , the elasticity can also be represented using Equation 3-3

$$\epsilon_i = \left|\frac{p_i}{Q_i}\frac{\partial Q_i}{\partial p}\right| \tag{3-3}$$

From (3-2) and (3-3), we have the Equation 3-4

$$p_i = mc + \frac{Q_i}{|\partial Q_i / \partial p|} \tag{3-4}$$

Subsequently, the hedonic function can be written as:

$$h(x_i) \equiv E[p_i|x_i] = E[mc|x_i] + E\left[\frac{Q_i}{|\partial Q_i / \partial p|}\middle| x_i\right] \tag{3-5}$$

This equation consists of both marginal cost (first term) and markup price (second term). The first term is also dependent on the customers' demand. The challenging question is how the first and the second terms interact with each other and how to calculate the market price.

Fortunately, we can use the regression analysis as an empirical tool to estimate the relationship between the response variable (cloud price) and explanatory variables (cloud characteristics). This is the basic idea of the hedonic approach. The idea of predicting hedonic price has been consolidated by Haas, Court, and Waugh and theorized by Lancaster [141] and Rosen [142]. According to Brachinger [143], the functional relationship of hedonic prices can be defined as:

$$MWTP = \frac{\partial p}{\partial x_i}(x) = \frac{\partial h}{\partial x_i}(x), (i = 1 \cdots k) \tag{3-6}$$

where "MWTP" is the marginal willingness to pay, "p(x)" is the price function, "h(x)" is the hedonic function. x_i is the characteristic of a product. Practically, there are four common types of hedonic forms (linear, semi-log, log-log, or Cobb-Douglas and logarithmic, see Table 3—3). But, as both Rosen and Halvorsen et al. [146] indicated that "The appropriate functional form for the hedonic equation cannot, in general, be specified on theoretical grounds." This means that a practical solution to select a particular function form is really dependent on a dataset in hand, which is to examine which function form to be goodness-of-fit with a collected dataset. Halvorsen proposed a statistical procedure to select a functional form with a Box-Cox methodology that is basically to use the likelihood ratio to examine the appropriateness of the alternative functional forms. However, Cassel et al. [147] argued that Box-Cox transformation is inadequate for the purpose of predicting hedonic prices because:

1. It is not necessary to increase the accuracy of price prediction. In fact, it could lead to a poorly estimated result, which had been demonstrated by Rao [136] .
2. The collected data may contain some negative values, but the traditional Box-Cos function does not allow any negative values because any negative number raised to non-integer real power would become imaginary.

3. Because the mean predicted, the value of the untransformed dependent variable is not necessary to be equal to the estimated mean that has been transformed. As a result, the nonlinear transformation will introduce a bias for the untransformed variable.

Overall, the nonlinear transformation results would be challenging to be explained.

Table 3—3 Common Regression Function Forms for Hedonic Analysis

Function form	**Hedonic Regression Equations**	**The inverse of Price Elasticity (Lerner Index)**	**Hedonic Price**	**Interpretation of Hedonic price**
Linear	$p = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$	$1/\epsilon^D = \beta_i \frac{x_i}{p}$	$\frac{\partial p}{\partial x_i} = \beta_i$	Marginal change
Quadratic	$p = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{m} \gamma_i x_i^2$	$1/\epsilon^D = (\beta_i + 2\gamma_i x_i)\frac{x_i}{p}$	$\frac{\partial p}{\partial x_i} = \beta_i + 2\gamma_i x_i$	linear marginal change
Cubic	$p = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{m} \gamma_{i+1} x_i^2 + \sum_{i=1}^{n} \theta_i x_i^3$	$1/\epsilon^D = (\beta_i + 2\gamma_i x_i + 3\theta_i x_i^2)\frac{x_i}{p}$	$\frac{\partial p}{\partial x_i} = \beta_i + 2\gamma_i x_i + 3\theta_i x_i^2$	Quadratic marginal change
Linear Intrinsic & Extrinsic & Time Dummy	$p_{ij}^t = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=1}^{l} \xi_j z_j + \sum_{t=1}^{T} \delta_t D_t$	$1/\epsilon^D = (\beta_i x_i + \xi_j z_j + \delta_t D_t)/p_{ij}^t$	$\frac{\partial p}{\partial x_i} = \beta_i + \alpha_{jt}$	Marginal change + multiple fixed effects
Exponential or Semi-log	$p = \beta_0 \prod_{i=1}^{k} e^{\beta_i x_i}$, or $lnp = ln\beta_0 + \sum_{i=1}^{k} \beta_i x_i$	$1/\epsilon^D = \beta_i x_i$	$\frac{\partial p}{\partial x_i} = \beta_i p$	Growth Rate
Semi-log + Dummy Variable	$lnp^t = \alpha_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{t=1}^{T} \delta_t D_t$	$1/\epsilon^D = \beta_i x_{ji} + \delta_t D_t$	$\frac{\partial p}{\partial x_i} = \beta_i p^t + \alpha_t$	Power of marginal change + time fixed the effect
Power or Double log	$p = \beta_0 \prod_{i=1}^{k} x_i^{\beta_i}$ or $lnp = ln\beta_0 + \sum_{i=1}^{k} \beta_i lnx_i$	$1/\epsilon^D = \beta_i$	$\frac{\partial p_i}{\partial x_i} = \frac{\beta_i}{x_i} p$	Partial Elasticities
Logarithmic	$p = \beta_0 + \sum_{i=1}^{k} \beta_i lnx_i$	$1/\epsilon^D = \frac{\beta_i}{p}$	$\frac{\partial p}{\partial x_i} = \frac{\beta_i}{x_i}$	Marginal change of logarithmic

3.3.3 Hedonic Model for Cloud Price

To the best of our knowledge, only limited studies of hedonic analysis had been conducted for cloud pricing, although the hedonic model has been widely applied in other industries, such as real estate, automobile, hotel, airline, and recreation. El Kihal et al. [84] were among the first

presented a simple hedonic analysis regarding Infrastructure as a Service (IaaS) clouds. The result of the hedonic analysis is not compelling because the adjusted R-squared was 43% (IBM). Nevertheless, they initiated the hedonic model for further study of cloud prices. Mitropoulou et al. [85] [162] provide a hedonic price index for cloud price comparison purposes among 23 CSPs.

In summary, previous studies left a large gap of hedonic modeling for the cloud pricing in terms of exploring different alternative hedonic forms, reducing regression errors, increasing R-squared values and adding practical values for cloud decision-makers. In this chapter, we show how to overcome many of these issues.

3.4 Hedonic Function for Cloud Pricing

3.4.1 Hedonic Function

By the extension of previous research for cloud prices, we first define the simplest hedonic function form of linear regression using OLS (Ordinary Least Square) method for our initial test. It can be directly interpreted as the mean coefficient values multiplied by independent variables (a vector of cloud characteristics) plus an error term:

$$p(X) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon \tag{3-7}$$

where $X = (x_1, x_2, .., x_k)$, x_i are independent variables and also a vector to represent different cloud characteristics, such as RAM, CPU core, virtual CPU, storage size and network bandwidth. The "*k*" is the number of cloud characteristics. "*p*" is a dependent variable to represent cloud instance price, which can be observed from CSP's web price catalog. Both independent and dependent variables are numerical values. β_i is the linear coefficient and β_0 is the interception point of the linear equation and ε is the error term or noise. The issues of the linear model are:

1) It may create substantial errors because of underfitting. The previous analytic results [84] demonstrated the R-squared value could be as lower as 46%.
2) This model cannot capture the price change due to time variation for the unbalanced panel data. In other words, it is impossible to measure the price change along with the temporal domain.
3) This model also ignored extrinsic features.

4) Moreover, some of the cloud characteristics provided by each individual CSP, such as a dedicated server, burstable CPU, and OpenStack API, cannot be captured due to the binary nature of these features. Therefore, it could lead to inaccurate pricing estimation.

In order to overcome these issues, we have to develop many sophisticated hedonic function forms to minimize the regression errors based on the collected datasets.

3.4.2 New Hedonic Function Form

One of the solutions to minimize regression error due to time dependency is to add another independent variable for the unbalanced panel data, namely time dummy or indicator variables to the OLS equation. This variable can capture the chronological influence of cloud prices. As a result, Equation 3-8 would become Equation 3-9 as:

$$p(\boldsymbol{X}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=1}^{l} \xi_j z_j + \sum_{t=0}^{T} \delta_t d_t + \varepsilon \tag{3-8}$$

$$\boldsymbol{X} = \langle x_1 \cdots x_k, z_1 \cdots z_j, d_1 \cdots d_T \rangle$$

$$z_j \in \{0,1\},\ \ d_t \in \{0,1\}, \sum_{t=0}^{T} d_t = 1$$

Here, d_t is the time dummy variable. Often, the unit of T is the number of years. δ_t is the coefficient value. ε is the error term that generates by both numerical and binary variables.

Furthermore, in order to capture the categorical variable of cloud service characteristics, we separate all cloud characteristics into two categories, namely intrinsic and extrinsic characteristics. The intrinsic characteristics are closely associated with cloud infrastructure costs. They often appear to be continuous variables. In contrast, the extrinsic characteristics are the binary variable. It means that CSPs can either support or not for a particular cloud instance. These service features will only add values to the customers when some intrinsic cloud characteristics are enabled. Let alone they often have no instrumental values to customers. Subsequently, we can develop further Equation 3-9 to be as follows:

$$p(\boldsymbol{X}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=1}^{l} \xi_j z_j + \sum_{t=0}^{T} \delta_t d_t + \varepsilon, \tag{3-9}$$

$$\boldsymbol{X} = \langle x_1 \cdots x_k, z_1 \cdots z_j, d_1 \cdots d_T \rangle$$

$$z_j \in \{0,1\},\ \ d_t \in \{0,1\}, \sum_{t=0}^{T} d_t = 1$$

where $z_j$is the binary variable (In general, z_j can be a categorical variable.) that represents extrinsic cloud characteristics j and "*l*" is the number of the extrinsic characteristics. ξ_j is the coefficient of the binary variable. ε is the term of combination errors for both intrinsic and extrinsic characteristics plus time dummy variable. If we take the derivative of Equation 3-9, we should have a vector of derivatives.

$$\nabla p(X) = [\beta_1 \beta_2 \cdots \beta_k \ \zeta_1 \zeta_2 \cdots \zeta_l \ d_1 d_2 \cdots d_T \,]^T \tag{3-10}$$

Intuitively, the extrinsic cloud characteristics are similar to spatial fixed effects in the property data application. Kuminoff et al. [154] suggested adopting a combination of spatial fixed effects, quasi-experimental identification, and temporal controls would provide an unbiased result because of many unobserved characteristics. If all characteristics are explicit, Cropper et al. [155]suggested that linear and quadratic Box-Cox forms would produce the best results.

However, Triplett [148][149] Griliches [150] and Gordon [152] indicated that the semi-log form has frequently emerged as "best" in hedonic function form tests. As a result, we can rewrite Equation 3-9 as a semi-log form. It can handle the substantial price variation of cloud instances for a long time period.

$$\ln [p(X)] = \beta_0 + \sum_{i=1}^{k} \beta_i \boldsymbol{x_i} + \sum_{j=1}^{l} \xi_j \boldsymbol{z_j} + \sum_{t=1}^{T} \delta_t \boldsymbol{d_t} + \varepsilon \tag{3-11}$$

$$\boldsymbol{X} = \langle x_1 \cdots x_k, z_1 \cdots z_j, d_1 \cdots d_T \rangle$$

$$z_j \in \{0,1\},\ \ d_t \in \{0,1\}, \sum_{t=0}^{T} d_t = 1$$

Transformations will make sense if the dataset has the following features [153]:

1. The variance of the errors is unequal or heteroscedasticity.
2. The ratio between max and min is greater than 5.
3. The scatterplot of dependent and independent variables is curved.

4. The data points are skewed, which the data has a long right tail.
5. All values are positive.

Generally, the transformation will consider the response variable (cloud instance price) first and then both the explanatory and response variables. Another solution to reduce the regression errors is to develop a polynomial regression formula, which is to add multiple high order terms for the independent variables if the collected dataset shows that the relationship between a dependent variable (cloud price) and independent variables (cloud service characteristics) is not linear.

This chapter considered a variety of hedonic function forms, as shown in Table 3—3, to minimize estimated errors

3.5 Performance Evaluation

3.5.1 Datasets and Assumptions

3.5.1.1 AWS Panel Data

The AWS panel data comes from two sources: 1) internet archive [144], 2) Amazon annual reports [145]. The data was recorded or sorted based on the time sequence that AWS released a new service catalog every time.

Although Amazon started its AWS business as early as in 2006, AWS had a limited number of characteristics for its cloud services. Most of them belonged to intrinsic characteristics. In fact, AWS did not offer cloud services to the general public until 2007. Consequently, the cut off time for the panel data test began in 2008. In the beginning, AWS offered only four instances to the public. Later, AWS gradually added more types of cloud instances to its service catalog. Each instance has a particular configuration, the Application Programming Interface (API) name, and its price tag. After 2013, AWS superseded some previous generation of Elastic Compute Cloud (EC2) and replaced it with a current generation of instances.

AWS pricing catalog is evolving from time to time due to the innovation of cloud technologies and pricing models. Some intrinsic variables are mixed with numerical and categorical values. Moreover, AWS sometimes changes its CPU measurement in response to the cloud market competition [156]. Therefore, we have made the following assumptions in order to simplify the AWS panel dataset:

1. For optimized instances, AWS uses HDD for d-serial instance and Non-Volatile Memory (NVMe) SSD for i3-serial instance. The rest of the instances are either SSD or EBS only. In order to simplify the calculation, we assume these different characteristics of instance storage to be the same as HDD in terms of unit cost.
    - However, the prices of SSD, NVMe SSD and HDD are different in the market. So, this assumption will contribute the certain price variations in our analysis.
2. The networking performance in the AWS catalog is mixed with numerical and categorical variables. As a result, we unified all variables with the same numerical unit, which the category of very low is equal to "1", "low" is equal to "2", "low to moderate" is equal to "3", "Moderate" is equal to "4", "high" is equal to "5" and "Up to 10 GBits" is equal to "6". This assumption might also create some errors because "1" might not be necessarily equivalent to 0.1 GBits links.
3. AWS has two different types of instance prices for two operation systems: Linux and Windows. For this chapter, we only use the Linux price on-demand. The price ratio of Linux and Windows is ranging between 1.00 and 2.05. It is dependent on the size or capacity of the instance. AWS provides customer long-term subscription discounts if cloud customers have a long-term commitment, which is the so-called "reserved price." This is another aspect of the problem that will be dealt with separately in other research.

3.5.1.2 Computer Hardware Data

In order to make a price comparison between cloud service (IaaS) and general computer hardware with the influence of Moore's law, we include the general computer hardware market data of CPU, GPU, SSD, flash memory, storage, Hard Disk Drive (HDD) [164]. There have been some other works [77][78][161] for cloud price comparison, but they only focused on the cloud compute or storage resources in isolation. Our study takes into account all the dependent variables.

3.5.1.3 Cross-Sectional Data

The cloud characteristics are released by different CSPs almost daily. Capturing all cloud characteristics is impossible. Due to the limitation of the dataset, we only have a total of 55 extrinsic cloud characteristics. Among them, 48 are considered to be the typical cloud characteristics, such as Pay-As-You-Go, Web interface, API, and Free-Transfer-In, in which nearly all CSPs provide these common characteristics for their service. As a result, they have become the baseline of extrinsic cloud characteristics. In this study, we limit the number of extrinsic characteristics for our analysis because some of the extrinsic characteristics are

insignificant (p-value > 0.05) such as vertical scaling without a reboot, OpenStack-compatible API and backup snapshot due to a limited number of data points. Furthermore, each CSP started the cloud business at a different time. Some of them just launched the cloud business recently.

3.5.1.4 All Cloud Instances of Five Leading CSPs

According to the latest Gartner's magic quadrant market report for the public cloud of IaaS [157] AWS, Microsoft and Google are the market leaders and Rackspace is one of the challenges and closely follows these three (see Table 3—4). Linode is one of the leading competitors in the US IaaS market.

Table 3—4 Five Leading Public Cloud Service Providers

Name of CSP	No of Instances prices	No of Baseline Characteristics	No of Host Domains (30-Jan-17)	No of Host Domains (30-Mar-17)
AWS	76	48	948,207	1,015,002
Microsoft Azure	69	48	142,854	149,175
Google Cloud Platform	21	48	599,846	630,117
Rackspace	19	48	504,624	487,827
Linode	14	48	210,106	220,717

Note that some of the extrinsic characteristics add extra costs for the cloud services, for example, a 10-node Hadoop cluster would have the extra cost of 0.15/per hour. In order to make a fair and horizontal comparison among different CSPs, we only track some extrinsic cloud characteristics across the board, which have no extra charge for an instance price. We assume CSPs do not charge an extra price for their baseline service configuration in their price catalog. These extrinsic characteristics of cloud service often have binary values, which are either 0 or 1.

3.5.2 Test Design, Roadmap and Results

We start with the 1st test that is designed to analyze the cloud instance price. We adopt the AWS cloud catalog dataset for the 2014 year (see Section 3.5.2.1). It is a simple OLS test. The purpose of this test is to examine the relationship between cloud instance prices (on-demand price for Linux OS) and its intrinsic characteristics. According to AWS, ECU (virtual server) resource is equivalent to CPU capacity of one 1.0-1.2GHz 2007 Opteron or 2007 Xeon processor. However, AWS has quietly adopted the unit of vCPU measurement in 2014. Each vCPU would correspond to a hyperthread of Intel Xeon core (clock speed) except t-serial instances. The purpose of a

hyperthread technology is to increase CPU performance by sharing the computational workload among multiple cores. The value of ECU usually is higher than vCPU except for t-serial instances.

The second test consists of time dummy variables based on AWS unbalanced panel dataset between 2008 and 2017 (as discussed in Section 3.5.2.2). This test is an extension of the OLS. However, we add the second and third-order polynomial terms into the linear equation in order to increase R- squared and reduce p- values.

The last test is to compare cloud instance prices among five different CSPs based on the cross-sectional dataset in 2017. This test is designed to add the extrinsic variables into the hedonic function form. It is to analyze the impact of cloud extrinsic characteristics on the price of baseline instance configuration (as discussed in Section 3.5.2.3). A roadmap of these three tests is illustrated in Figure 3—3, which illustrates how we demonstrate the cloud extrinsic characteristics. In doing so, we report some performance and decision parameters of the preliminary models and then the full cross-sectional data for the final model.

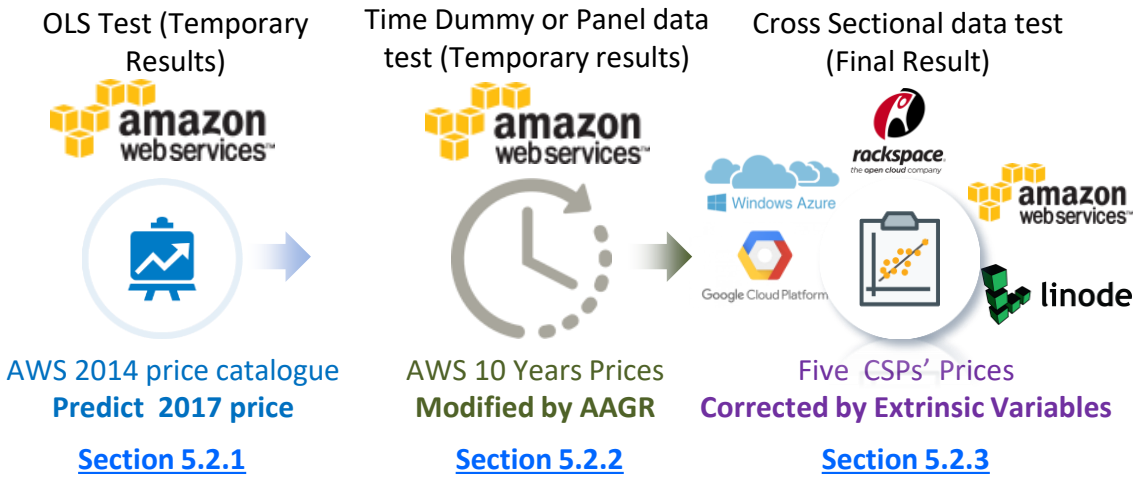


Figure 3—3 Simple Roadmap of Three Tests

This research used R and R Studio to implement both panel data and cross-sectional data regression analysis.

3.5.2.1 AWS Instance Price Test

According to our test design, we construct a simple linear regression model between Linux on-demand price and six explanatory intrinsic variables and then have a normality test and residual plots with instance price for the dataset.

Both R-squared and adjusted R-squared values are about 0.82-0.83 so that OLS only explains 82% of data points. Based on both the normality test and residual plots, we can see two outlier data points. These outlier points may cause regression errors. If we excluded these two points, the R-squared values could be increased.

We also notice that the coefficient of both vCPU and Bit (Architecture 32-bit or 64-bit) is negative. This may also be triggered by regression errors. By excluding the new large GPU instances or outlier data points, we can improve the residual values of this OLS dramatically. The R-squared values are lifted to about 93% (Table 3—5).

Table 3—5 The Linear Form of Hedonic Function for 2014

Coefficients	Estimated β	Std. Error	t-value	Pr(>|t|)
Intercept	**-0.3377**	**1.06E-01**	**-3.176**	**0.00186 ****
RAM	0.0049	3.98E-04	12.326	< 2e-16 ***
VCPU	0.0181	5.94E-03	3.044	0.00283 **
Storage	0.00005	8.42E-06	5.897	3.01e-08 ***
network performance	0.1755	2.67E-02	6.586	1.02e-09 ***

Residual standard error: 0.5949 on 130 degrees of freedom
Multiple R-squared: 0.9273, Adjusted R-squared: 0.9251
F-statistic: 414.8 on 4 and 130 DF, p-value: < 2.2e-16

Note: "*"means a significant code of p-value, "***"= p<0.001, "**"= p<0.01, "*"= p < 0.05

Furthermore, the *p*-values of ECU, CPU, and Bit become insignificant. The test has proved the Gartner's claim [156], which AWS quietly shifted from ECU to vCPU. Therefore, we can safely exclude ECU and CPU as independent variables with a limited impact on R-squared and adjusted R-squared values. By extracting hidden values from the intercept (or beta zero), we can transform it into a semi-log form and add polynomial higher-order terms into the OLS equation (as shown in Table 3—6).

Table 3—6 The Semi-log Form of Hedonic Function for 2014

Coefficients	Estimated β	Std. Error	t-value	Pr(>|t|)	EXP(βi)
Intercept	-5.04E+00	2.14E-01	-23.483	< 2e-16 ***	0.01
RAM	8.71E-03	1.64E-03	5.307	4.82e-07 ***	1.01
RAM^2	-1.02E-05	2.32E-06	-4.402	2.25e-05 ***	1.00
RAM^3	3.81E-09	8.07E-10	4.726	5.98e-06 ***	1.00
VCPU	7.87E-02	1.23E-02	6.414	2.55e-09 ***	1.08
VCPU^2	-6.83E-04	2.14E-04	-3.194	0.00177 **	1.00
storage	1.99E-05	7.13E-06	2.792	0.00605 **	1.00
network performance	1.28E+00	9.45E-02	13.571	< 2e-16 ***	3.60
network performance^2	-9.40E-02	7.29E-03	-12.882	< 2e-16 ***	0.91
EBS.O	-1.22E-04	4.18E-05	-2.913	0.00423 **	1.00

Residual standard error: 0.4793 on 125 degrees of freedom
Multiple R-squared: 0.9118, Adjusted R-squared: 0.9055
F-statistic: 143.6 on 9 and 125 DF, p-value: < 2.2e-16

One issue with the linear form is that the absolute value of the intercept β_0 (-0.338) is the highest in comparison with other β_i (or hedonic) values. A practical interpretation of this negative β_0 is that AWS would pay customers upfront for on-demand instance, which is not the case. One of the reasons for the higher absolute β_0 value is there are other hidden variables within β_0. With the semi-log form, the β_0 value is down from 0.338 to 0.0064. Although both R- squared values slightly decline by about 2%, the β_0 value is reduced by nearly 53 folds. On the other hand, the result of the semi-log form is difficult to be interpreted because of the higher-order polynomial terms with the negative β values. The model becomes quite sensitive for the large instance configuration, especially for the characteristics of RAM and network. One of the reasons is that AWS may insert a volume discount mechanism for large instances. The other possible reason is AWS does not give the resource-level permission to reboot, start, delete, detach EBS volume, etc. for cloud customers to specify a resource in every instance action in order to maintain control of its cloud infrastructure resource pool.

As noted in AWS 2017 catalog, AWS offers a wide variety of configurations for its computing instances such as cc1.4xlarge (cluster compute quadruple extra-large VM), cg1.4xlarge (GPU VM), and m1.small (general-purpose small resource VM). To predict a cloud price of an average configuration resource in the AWS 2017 catalog, we used the m4.10xlarge instance, which is one of the general-purpose instances and provides a balance of computing memory and network resources. It is designed to support different computing environments such as web applications

or line of business or LoB (The letter "m" stands for "general purpose," "4.10" means the size of computing and network resources, "xlarge" stands for extra-large.) The detail configuration of this instance is RAM=160, ECU =124.5, vCPU=40, CPU=3.112, storage=0, Network Performance=10, EBS.O= 4000.

Based on this configuration, we can predict the price of the m4.10xlarge instance as $2.925 (linear form) or $2.961 (semi-log form). The real price for m4.10xlarge instance is $2.155 (see Table 3—7). Although this prediction value is within 95% of the confidence interval, the predicted fitted value is about 36% higher than the real price value, and the price range between low and upper bound is high, but the linear form is slightly better than the semi-log. This might be due to many factors, such as different function forms, sample size, and skew dataset. Moreover, we have not taken consideration of time impact. Based on Moore's law, the price of computer resources should decrease by about -50% per annum. This issue leads to our next topic of analysis, namely the time dummy variable.

Table 3—7 Predicting Price of a Cloud Instance with m4.10xlarge Configuration

With a 95% confidence interval	Fitted Value	Real price	Price difference Δp	Accuracy	Lower	Upper
Predicted Value by Linear form	$2.925	2.155	$0.77	64.3%	1.716	4.134
Predicted value by semi-log form	$2.961	2.155	$0.806	62.9%	1.110	7.898

3.5.2.2 AWS Panel Data Test with Time Dummy Variables

If we consider the time variables, the total number of data points (instances) of an unbalanced panel dataset is 837 between 2008 and 2017. The number of explanatory or intrinsic variables is almost identical either using vCPU or ECU. However, ECU is AWS long-term measurement for CPU resources. The time dummy variables are 9 (10 years, T-1 time dummy variable, Table 3—8). The linear Q-Q plot shows that it is highly skewed, but after a semi-log transformation of instance prices, the Q-Q plot appears to be much better (see Figure 3— 4).

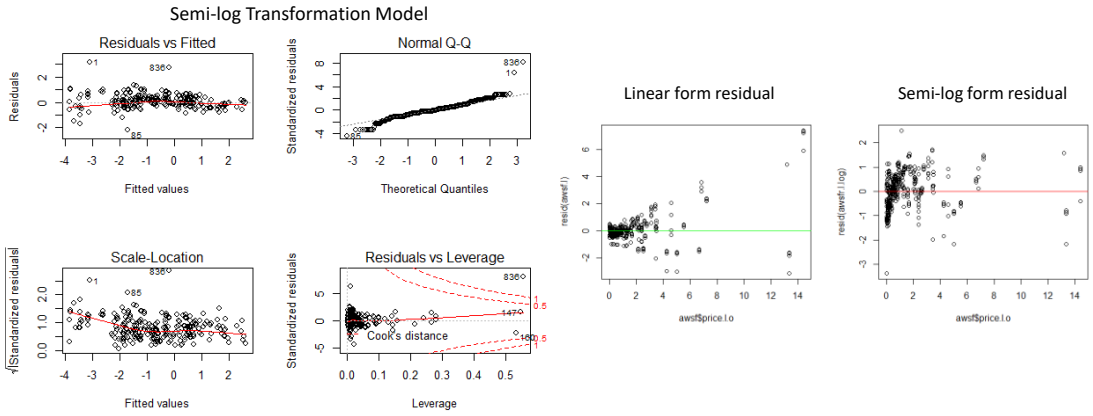


Figure 3—4 Log Transformation model and Residual Errors Plots 2008 - 2017

The primary objective of the semi-log transformation is for time dummy variables inference. The initial linear model test only shows 7 variables, including four-time dummy variables (2017, 2016, 2015 and 2014) are significant. It means that we can only infer for four years. The R-squared and adjusted R-squared values are 0.8271 and 0.8235, respectively.

If we take the semi-log transformation, more time dummy variables become highly significant. The R-squared and adjusted R-squared values drop slightly to 0.8148 and 0.8109, respectively. If we add high order polynomial terms into the semi-log form, the test result is promised (see Table 3—8). There are two additional considerations to transfer hedonic function from linear to semi-log form:

Table 3—8 AWS Panel Data Regression Test with Time Dummy Variables (2008-2017)

Coefficients	Estimated β,δ	Std. Error	t-value	Pr(>|t|)	Annual Rate A_t	CAGR C_t
intercept	-3.87E+00	1.42E-01	-27.167	< 2e-16 ***		
RAM	2.46E-03	1.99E-04	12.315	< 2e-16 ***		
ECU	4.48E-02	1.93E-03	23.287	< 2e-16 ***		
ECU^2	-2.47E-04	1.59E-05	-15.578	< 2e-16 ***		
ECU^3	3.29E-07	3.23E-08	10.202	< 2e-16 ***		
Storage	2.26E-05	2.77E-06	8.159	1.27e-15 ***		
Net Perf	6.08E-01	2.77E-06	9.983	< 2e-16 ***		
Net Perf ^2	-6.46E-02	7.01E-03	-9.218	< 2e-16 ***		
Net Perf ^3	1.95E-03	2.27E-04	8.592	< 2e-16 ***		
bit	4.95E-02	2.17E-03	2.17E-03	< 2e-16 ***		
d17	-2.70E+00	1.62E-01	-16.709	< 2e-16 ***	-1.49%	
d16	-2.69E+00	1.54E-01	17.479	< 2e-16 ***	0.00%	
d15	-2.69E+00	1.54E-01	-17.479	< 2e-16 ***	-9.15%	
d14	-2.59E+00	1.55E-01	-16.769	< 2e-16 ***	-44.07%	
d13	-2.01E+00	1.64E-01	-12.261	< 2e-16 ***	-77.71%	
d12	-5.08E-01	1.21E-01	-4.196	3.02e-05 ***	-14.10%	
d11	-3.56E-01	1.24E-01	-2.859	0.00435 **	-6.69%	
d10	-2.87E-01	1.25E-01	-2.366	0.01822 *	-18.54%	
d9	-8.16E-02	1.25E-01	-0.65	0.51563	-7.83%	
d8	0.00E+00	baseline	baseline	baseline		
AAGR					-20.0 %	
Compound Average Growth Rate						-25.9%

Residual standard error: 0.4913 on 817 degrees of freedom
Multiple R-squared: 0.902, Adjusted R-squared: 0.8999
F-statistic: 418 on 18 and 817 DF, p-value: < 2.2e-16

1. The price of cloud infrastructure is closely associated with computer hardware. According to Moore's law, the hardware price depreciation rate is exponential in the time domain.
2. Previous experiences [159], [160] suggested the adoption of the semi-log model if a test is designed for a longer-term comparison.

Based on the above test result with the time dummy, we can calculate Annual Growth Rate (A_t) and Average Annual Growth Rate (AAGR) by the following two equations:

$$A_t = \frac{\left(e^{\delta_t} - e^{\delta_{t-1}}\right)}{e^{\delta_{t-1}}} \tag{3-12}$$

$$AAGR = \left(\left(\sqrt[T-1]{\prod_{t=2}^{T}(1 + A_t)} \right) - 1 \right) \times 100 \quad (3\text{-}13)$$

Note:

1. β_0= - 0.021, It is a combination of all explanatory variables. The value appears to be close to zero. It is a good indicator.
2. The coefficient values $\delta_1$is relative to 2008. It is emerged into β_0 value. Subsequently, "t" starts from 2.
3. There were no price changes between 2015 and 2016 after a significant discount in 2013 and 2014.
4. We use the geometric mean method to compute AAGR for the years 2008 to 2017, the rate of depreciation is -19.98%≈ -20%

Overall, AWS AAGR or price reduction rate is far less than what Moore's law prediction [158], which is about 50% per annum in general. The gap between AWS AAGR and Moore's law prediction is 50%- 20 %=30%. To a certain extent, this price gap indicates why cloud customers are willing to pay more than the benchmark price of computer hardware (see Figure 3—5). We also see that AWS made a substantial price discount in 2013 and 2014. It may indicate a seven-year life cycle of computer assets if we consider AWS bought its cloud hardware assets in 2006. This is actually in align with Walker's [77] conclusion.

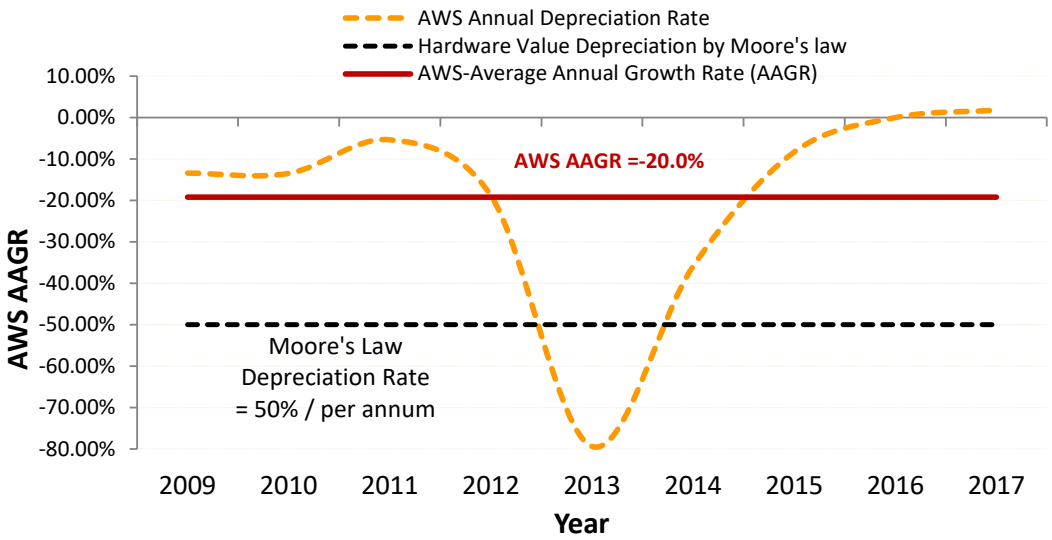


Figure 3—5 Comparison of AWS Cloud Depreciation Rate Vs. Moore's Law

The logic for this comparison can be justified by the following reasoning if we assume that the cloud instance configuration is ceteris paribus. Moreover, we assume that the independent time dummy variables do not interact with other independent variables; then, we should have the following two equations:

$$\ln((1+r)^t p) = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \delta_t d_t + \varepsilon \tag{3-14}$$

$$\ln((1+r)^{t-1} p) = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \delta_{t-1} d_{t-1} + \varepsilon \tag{3-15}$$

where, r = depreciation rate. Subtract (14) with (15) we should have the following equation (3-16).

$$\begin{gathered} r + 1 = e^{\delta_t - \delta_{t-1}}, \\ r = \frac{e^{\delta_t} - e^{\delta_{t-1}}}{e^{\delta_{t-1}}}, \quad r = A_t \end{gathered} \tag{3-16}$$

Based on the proof, we can derivate our conclusion that the AWS AAGR is around -20.0%/per annum in comparison with Moore's law.

By taking consideration of the impact of the time-dummy variable, the predicted price can be further updated. Alternatively, we can also use Compound Average Growth Rate (CAGR) to estimate the time impact, which is approximately close to AAGR. The CAGR formula is:

$$CAGR = C_t = \left(\frac{V_e}{V_s}\right)^{\frac{1}{T-1}} - 1 \tag{3-17}$$

where $C_t$is the compound average growth rate, V_e is the end value of the time period of "T" and V_s is the start value of the time period. Using the above prediction price in Table 3—7 as an example, we can correct the prediction result with CAGR in the following formula.

$$P_f = P_p \times (1 + A_t)^{(t_f - t_p)} \tag{3-18}$$

where P_f is for the future price and P_p is the present price, t_f is the future year value and $t_p$is the present year.

In order to predict future prices accurately, we have to exclude the future year from our dataset when we calculate AAGR. In our case, it is 2017 data points. Subsequently, the value of $AAGR_{2008-2016} \approx -17\%$.

If we use this $AAGR_{2008-2016}$ to predict the instance price of m4.10xlarge in 2017 based on the 2014 price catalog, we should have the following result (see Table 3—9) and the price difference between the real price and the predicted price (Δp) becomes negative.

Table 3—9 Estimate AWS Instance Price by Leveraging Time Dummy Variable

Within a 95% confidence interval	Fitted Value	Real Price	Price difference Δp	Accuracy	Lower	Upper
Predicted value (semi-log)	$1.693	2.155	-$0.46	78.59%	0.635	4.516

Now, the question is why consumers should be willing to pay more than the predicted price. The possible answer is the non-market characteristics of cloud services. From a CSP perspective, it is a part of CSP's marketing strategy to lead the cloud market. The common term is product or service differentiation. It leads to our next topic – a cross-sectional dataset test, which is to examine the cloud instance price that is contributed by extrinsic variables.

3.5.2.3 Cross-Sectional Data Test

Based on the five CSPs' product catalogs, we constructed a dataset that consists of the entire 199 cloud instances. The initial linear model shows that R-squared and adjusted R-squared values are about 0.8077 0.7885, respectively and the Q-Q plot shows the data is highly skewed. According to the above five principles of transformation (discussed in Section 3.4.2), we can transfer it onto a semi-log form. Once the transformation is done, the Q-Q plot shows a better result (see Figure 3—6) in comparison with the linear form.

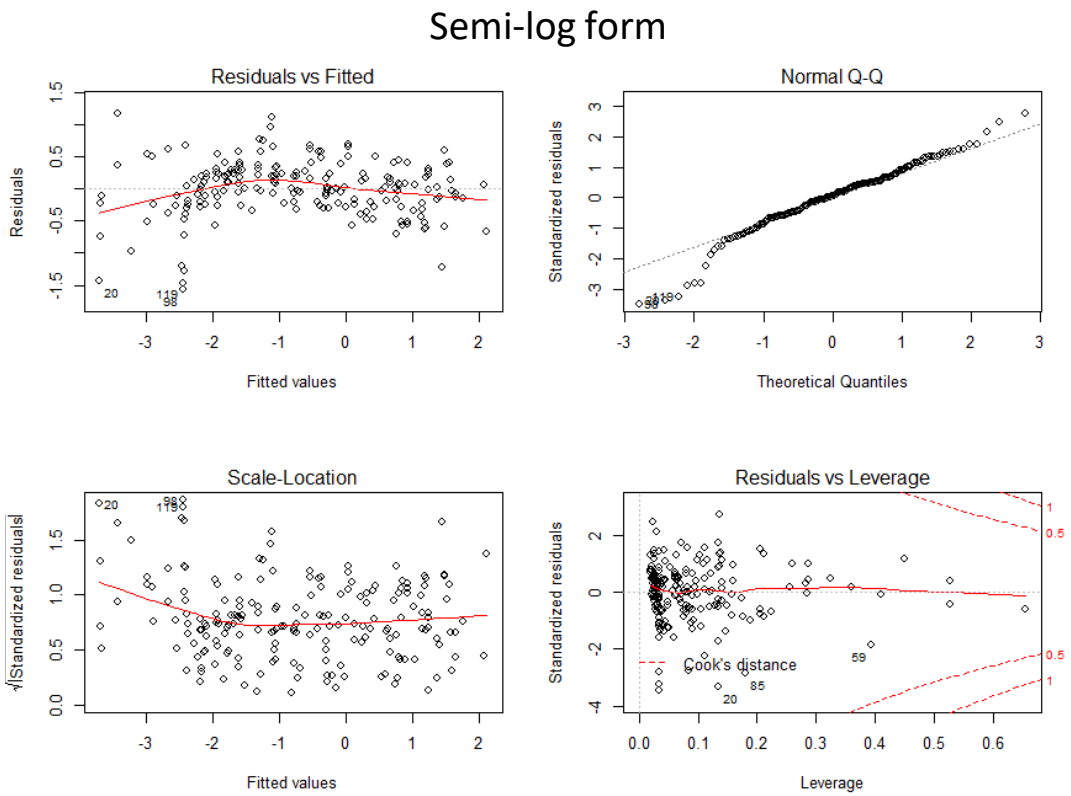


Figure 3—6 Semi-log transformation Form

By a combination of semi-log transformation, adding a high order of polynomial terms and excluding just a few highly outlier points, the R-squared and adjusted R-squared values are increased more than 10% up to 0.913 and 0.904 respectively (see Table 3—10). A discussion of these elements is noted below:

Table 3—10 Cross-Section Data Analysis Results with the Semi-log Transformation

Coefficients	Estimated β, ξ	Std. Error	t-value	Pr(>|t|)	EXP(Z)	Above the baseline
Intercept : $\boldsymbol{\beta_0}$	-2.68E+00	9.29E-02	-28.806	< 2e-16 ***		
RAM	2.50E-02	3.08E-03	8.122	8.14e-14 ***		
RAM^2	-1.14E-04	1.73E-05	-6.56	5.92e-10 ***		
RAM^3	1.47E-07	2.51E-08	5.875	2.10e-08 ***		
VCPU	1.98E-01	2.43E-02	8.154	6.72e-14 ***		
VCPU^2	-6.03E-03	9.76E-04	-6.18	4.41e-09 ***		
VCP^3	5.55E-05	1.02E-05	5.458	1.64e-07 ***		
Storage	2.68E-04	6.57E-05	4.073	7.03e-05 ***		
Storage^2	-2.53E-08	7.13E-09	-3.551	0.000494 ***		
Storage^3	4.20E-13	1.22E-13	3.428	0.000758 ***		
Network Performance	1.77E-05	2.70E-04	2.479	0.014257 *		
Arch	-7.67E-02	1.13E-02	-6.786	1.73e-10 ***		
Arch^2	1.19E-03	1.75E-04	6.826	1.40e-10 ***		
Free Transfer to dedicated location ξ_2	5.49E-01	1.68E-01	3.275	0.001292 **	1.732	73.2%
GPU instance ξ_3	3.43E-01	1.63E-01	2.104	0.036771 *	1.409	40.9%
Burstable CPU ξ_4	5.33E-01	2.19E-01	2.44	0.015692 *	1.704	70.4%
Dedicated serversξ_5	3.20E-01	1.29E-01	2.48	0.014097 *	1.377	37.7%
One account for all locationsξ_6	2.45E-03	3.33E-04	7.342	9.66e-12 ***	1.002	0.2%
Data Center Global Foot Print (AUS)ξ_7	2.82E-01	1.29E-01	2.19	0.029850 *	1.326	32.6%
Collocation ξ_8	4.00E-01	1.25E-01	3.194	0.001666 **	1.491	49.1%
48 Baseline Extrinsic Characteristics	0	Baseline	Baseline	Baseline	1	-
Average Extrinsic Price Value (AEPV)						43.4%

Residual standard error: 0.4607 on 173 degrees of freedom
Multiple R-squared: 0.9128, Adjusted R-squared: 0.9042
F-statistic: 106.5 on 17 and 173 DF, p-value: < 2.2e-16

1.) Our analysis selected 5 intrinsic variables for cross-sectional data. Some intrinsic variables, such as a storage feature of Enterprise Block Store (EBS) optimized excluded from this test because it is insignificant for the regression analysis.

2.) Based on the available dataset, we can make inference for 7 extrinsic variables (p-value is less than 0.05) with respect to a baseline characteristics of instance configuration (including, API, PAYG, Web interface, auto-scaling, resource usage monitoring, free transfer in, Free IP, load balancing, firewall, backup storage, credit card payment, volume discounts, free entry-level service and etc.).

3.) The value of ζ_1 that represents the baseline characteristics have been emerged into the β_0 value. Different baseline configurations will result in different β_0 values. It is dependent on the cross-sectional dataset. Ideally, the β_0 the value should be zero. However, it can only approach zero in reality.
4.) Dedicated servers can be considered as extra resources.
5.) Similar, free transfer to a dedicated location will give cloud customer mobility.
6.) Burstable CPU can save the CPU price. If you do not use your specified capacity, CSP will give you credit so that you can withdraw when you need it.
7.) The price of GPU Instance is much higher than the baseline instance with the configuration of the Intel CPU. AWS, GCP, and Azure provide the option of NVIDIA Tesla K80 GPU (launch price $3,169/per unit in 2017. In comparison with Intel Xeon E5-2673 V3 2.4-GHz chip, it costs $700/ per unit Jul 2017).
8.) The value ξ_6 of one account for allocation is minimal. It is basically submerged into the baseline characteristics, in which all CSPs provide this feature without extra cost.
9.) As Griliches indicated, the resulting regression is sometimes unstable. It could be varied, along with different circumstances. In the above case, if we change the configuration of the baseline extrinsic characteristics, the result will be entirely different.

Table 3—11 Predicted Price Including Extrinsic Values

With a 95% confidence interval	Fitted Value	Real price	Price difference Δp	Accuracy	Lower	Upper
Predicted Price with Ave Extrinsic Value	$2.428	2.155	$0.273	87.32%	0.911	6.476
Predicted Price with particular cloud extrinsic characteristic	$2.245	2.155	$0.09	95.81%	0.842	5.988

Now, we can answer the question that is raised before: "why cloud consumers are willing to pay nearly more than the predicted price." If we use Table 3—9 to further revise our price prediction by taking consideration of cloud extrinsic variables, we can find the predicted price is very close to the real price (see Table 3—11).

3.5.2.4 Predict Cloud Price for Different Instances

Notice that we can generalize Equation 3-18 as Equation 3-19 for future price prediction.

$$\hat{P}_f(X,Y) = (1+AAGR)^{Y-Y_0} \times \left(\beta_0 + \sum_{i=1}^{k}\beta_i x_i + \sum_{j=1}^{l}\xi_j z_j\right) \pm 1.96\sqrt{\frac{\hat{P}_f(X,Y)(1-\hat{P}_f(X,Y))}{n}} \quad (3\text{-}19)$$

where $X = \langle x_1 \cdots x_k, z_1 \cdots z_l, Y\rangle$, Y= future year, Y_0 = current year or present year, *n*= size of population for a dataset. (We adopt 95% Wald confidence intervals or first approximation). Furthermore, if we take the semi-log form, the equation can be presented as follows:

$$ln\hat{P}_f(X,Y) = (Y-Y_0) \times ln(1+AAGR) + \left(\beta_0 + \sum_{i=1}^{k}\beta_i x_i + \sum_{j=1}^{l}\xi_j z_j\right) \pm 1.96\sqrt{\frac{\hat{P}_f(X,Y)(1-\hat{P}_f(X,Y))}{n}} \quad (3\text{-}20)$$

We use this equation to estimate the future price of different cloud instances. The comparison of different AWS cloud instances produces the following prediction results (shown in Table 3—12).

We highlight three points for the prediction results:

1) The predicted prices usually are less than the real price. It means that AWS holds the price reduction pace due to its extrinsic values of a cloud instance.
2) For the standard instance, the predicted accuracy is approximately higher than 70% without consideration of extrinsic characteristics. (With one CSP, the extrinsic value cannot be compared)
3) For the latest generation cluster, the prediction accuracy is below 70%. It might be due to more extrinsic values that AWS has built into its price catalog.

Table 3—12 Predicted AWS Cloud Prices with Different Instance

Instance types	API name	Fitted Value with 95% CI	Real price	Price difference Δp	Accuracy	Lower	Upper
Standard	m1.medium	0.0842	$0.120	-0.036	70.2%	0.031	0.225
Standard	m4.10xlarge	$1.693	$2.155	-$0.462	78.6%	0.635	4.516
3rd Gen. Cluster	i3.8xlarge	$1.69	$2.496	-$0.808	67.6%	0.553	4.85
4th Gen. Cluster	c4.8xlarge	$0.979	$1.591	-$0.612	61.6%	0.252	3.197

Overall, once the predicted cloud price emerges, it can underpin the CIO to make the right strategic investment decision for IT infrastructure. Of course, he or she has to take consideration of other factors, such as business risks, workload growth, and volume discount and workload portability issues (or cloud vendor lock-in syndrome: "free to come and pay to leave").

3.6 Analysis and Discussion

We have illustrated how to use the hedonic analysis to predict the cloud instance price. From the unbalanced panel data, we can calculate the AWS' AAGR is approximate -20.0% per annum. Statistically, the time dummy variable is the same as a fixed effect. The net effect is the hedonic function to be shifted downwards (see Figure 3—7)

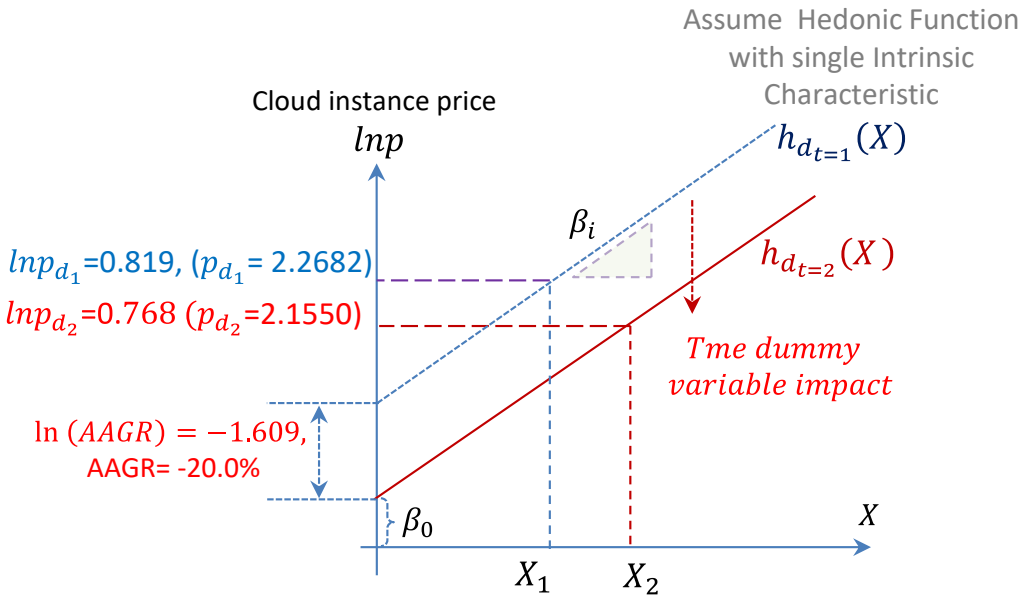


Figure 3—7 Impact of Time Dummy Variable on AWS Cloud Instance Price

In comparison with Moore's law prediction, the AWS price change rate (deflation) is at a much less slow pace than what Moore's law has predicted (-50% per annum). The reason that AWS can move beyond the competition just on price is its extrinsic characteristics that AWS can differentiate its cloud service from its competitors. AWS has developed more than 1,000 different cloud characteristics or features since 2006. Although we would not be able to analyze all extrinsic characteristics here, we can highlight some of the extrinsic characteristics among 5 leading CSPs (shown in Figure 3—8). The characteristic of a GPU instance is about 40.9% of cloud extrinsic value and data center global footprint (Australia) is 32.6% in comparison with the baseline configuration.

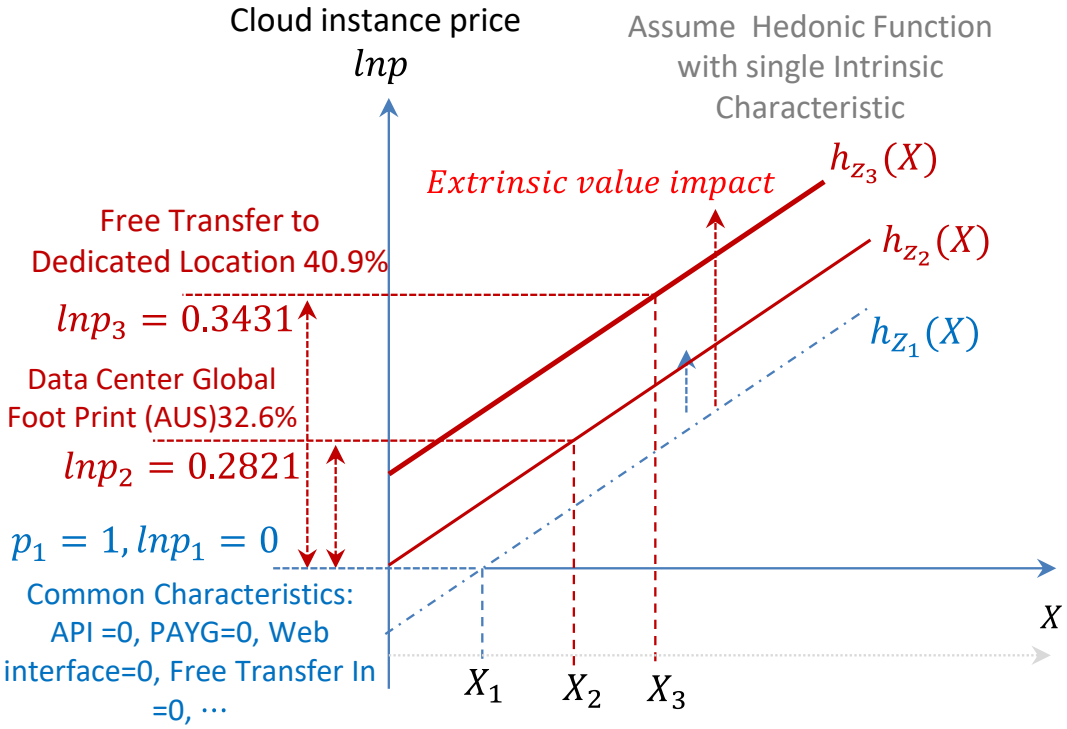


Figure 3—8 Impact of Extrinsic Variables on AWS Cloud Instance Price

Ultimately, the impact of the extrinsic variable is similar to the time dummy variable (or fixed effect). It only shifts the hedonic function either up or down. This means in order to avoid an estimated bias we should include the required cloud characteristics not only intrinsic variables

If we just compare the cloud instance prices based on the intrinsic variables alone (for standard configuration), AWS price is not the cheapest in comparison with the top 30 global leading CSPs. Its price is just slightly above the median one (The market median price is $146 marked as a notch.

AWS instance price is $149 is marked by a dashed line shown in a boxplot, Figure 3—9). However, AWS can still maintain over 31% of IaaS global market share and keep double digits revenue growth year on year (YoY). This is mainly due to the contribution of AWS extrinsic values of its cloud services, which cloud customers are willing to pay for.

In this chapter, we introduce the new concept of intrinsic and extrinsic variables that have been applied to the hedonic analysis of cloud pricing model. Moreover, we have mathematically proved that the time dummy or AAGR is equivalent to Moore's law impact if ceteris paribus. The AAGR plays a vital role in cloud price prediction.

In contrast to the previous studies that ignored the extrinsic variables impact on the cloud prices, we have clearly demonstrated that many extrinsic variables have significant values or fixed effect on the cloud price. The effective combination (or bundle) of intrinsic and extrinsic values does not only allow CSPs to slow the price reduction pace but also underpin the cloud market leadership.

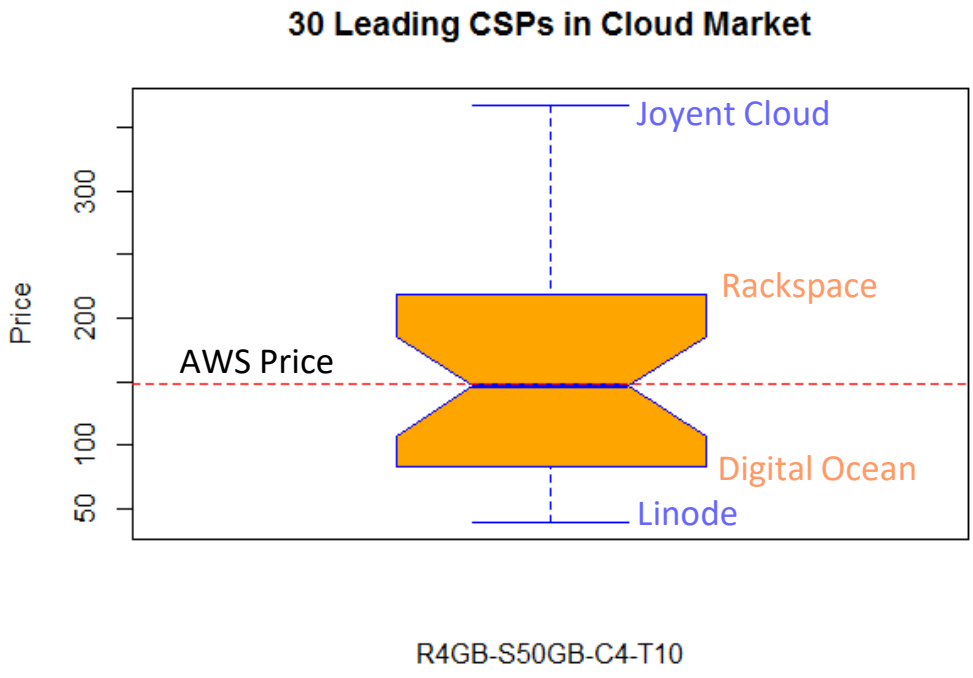


Figure 3—9 Box plot of 30 CSPs

Generally, the hedonic analysis is a practical or empirical approach to disclose the latent values of what customers are willing to pay for the quality changes. Ultimately, this research is to

leverage the hedonic concept to discover homogeneous cloud pricing patterns that are closely associated with heterogeneous cloud service characteristics, which are often hidden behind the complicated cloud pricing structure.

Our novel approach enables cloud customers to predict cloud service prices accurately based on their business application needs rather than purely on the cost of IaaS comparison. It means that cloud consumers can avoid many pricing estimation biases.

Another important implication is that it allows many CSPs to establish the correct performance benchmark based on the real value proposition of cloud services to compete with their market leader, not only just on the price.

3.7 Summary

This chapter mainly focuses on themes of cloud pricing for the new features or characteristics. The idea to differentiate the new cloud features and baseline service is to introduce the new concept of intrinsic and extrinsic values, which is inspired by the English Philosopher, G.E. Moore's influential work of Principia Ethica.

By proposing the hedonic analysis, the chapter exhibits that the hedonic pricing model can extract the many implicit cloud values based on the cloud customers' willingness to pay (W2P). This chapter demonstrated these implicit cloud values had become one of the competitive advantages for CSPs to lead the cloud market and increased the profit margin. The chapter shows that the model can capture the cloud value of non-marketable price, which is about 43.4% on average above the baseline. Unfortunately, this value is often ignored by many traditional cloud pricing models.

In addition, this chapter provides that the Average Annual Growth Rate (AAGR) of Amazon Web Services' (AWS) is about -20.0% per annum between 2008 and 2017, ceteris paribus. In comparison with Moore's law (-50% per annum), this value is at a far slower pace. The chapter argues this value is Moore's law equivalent in the cloud.

Chapter 4

Cloud Computing Market Segmentation

The topics of cloud pricing models and resource management have been receiving enormous attention recently. However, very few studies have considered the importance of cloud market segmentation. Moreover, there is no better, practical and quantifiable solution for cloud service providers (CSP) to the segment cloud market. This chapter proposes a novel solution that combines both hierarchical clustering and time series forecasting on the basis of the classical theory of market segmentation. In comparison with some traditional approaches, such as nested, analytic, Delphi, and strategy-based approaches, this method is much more effective, flexible, measurable and practical for CSPs to implement their cloud market strategies by rolling out different pricing models. The experimental results and empirical analysis show that this solution can efficiently segment cloud markets and also predict the market demands. The primary goal of this chapter is to offer a new solution so that CSPs can tail its limited cloud resources for its targeted market or cloud customers

4.1 Introduction

The issue of cloud pricing models, revenue, and resources management (cloud economics) is one of the most critical topics in the cloud computing [54] [55] because it does not only become increasingly important for many CSPs to implement their cloud business strategy but also allow them to innovate their business processes and models [166]. However, many previous studies [55] [165] only focus on finding an optimal solution from a pure CSP perspective (internal rationality) and often ignore market impacts (external rationality). In this study, we concentrate on the problem of cloud market segmentation, especially for business to business (B2B) markets by taking into account both CSP's resources and market factors[172].

This Chapter is derived from

- **Caesar Wu**, Rajkumar Buyya and Kotagiri Ramamohanarao, Cloud Computing Market Segmentation, Proceedings of the 13th International Conference on Software Technologies (ICSOFT 2018), ISBN: 978-989-758-320-9, Porto, Portugal, July 26-28, 2018

The B2B cloud market segmentation is believed to be a complex problem for many CSPs [187]. It is challenging because it involves many disciplines such as managerial decisions, market theory, price theory, cloud computing, and microeconomics. Moreover, it is often very subjective and arbitrarily.

We restrict our current study to B2B because the B2B market is more significant than a business to consumers (B2C) and consumer to consumer (C2C), according to the US Census Bureau [167]. Statista reported [196] the size of the Global B2B e-commerce market ($7.7 Trillion) is about 235% larger than B2C ($2.3 Trillion) in 2017. The cloud is a type of e-commerce as it shares the characteristic of online access [197]. Although the size of the B2B market is considerably large and it is crucial for CSP's business strategy and pricing, as of now, to the best of our knowledge, no work has been done on this topic. Yet, many CSPs urgently need to understand how to serve their targeted customers well for limited resources. Hence, our goal is to find a better solution to the segment cloud market.

To motivate the problem, we consider the following scenario. Suppose a local Internet Service Provider (ISP) has decided to expand its hosting business into the B2B cloud market with a limited investment budget. The CEO asks the management team to formulate a business strategy with different pricing models to grow both the cloud business revenue and profit. One of the most straightforward solutions is the "one-size fits all" or uniform pricing. It means that the ISP can set up a markup price for its desired profit margin while the customers have to decide either "take or leave it" regardless of what the customer's needs are. The subsequent question is, would this business strategy work. If not, what is an alternative solution that can be pursued?

An intuitive answer could be to deliver cloud services or products with personalized pricing to suit each customer's needs. However, it is impracticable for a CSP to offer personalized service and price because of the limited budget or resources. Fortunately, many individual customers have similar requirements, and their usage patterns may have some common characteristics, such as the size of computing (CPUs) and memory. It means that we can group these customers' demands. This idea leads to group pricing, which is also called market segmentation. The original concept of the market segmentation was introduced by Smith [166]. He defined the term of the segmentation at a strategic level, which "is based upon developments on the demand side of the market and represents a rational and more precise adjustment of product and marketing effort to

consumer or user requirements." He argued that proper market segmentation would lead to a successful business strategy.

In fact, the uniform pricing and personalized pricing are two extreme ends of the group pricing (Figure 4—1). Belleflamme et al. [167] stated: "The better the information about consumers, the finer the partition of the consumers into groups and the larger the possibilities for firms to extract consumer surplus."

Therefore, the goal of the segmentation process is to extract customer information, such as usage patterns or behaviors and then to develop various pricing models and service configurations to meet their needs. In fact, Yankelovich et al. [169] argued the proper market segmentation should meet the following criteria:

1) Align with the company's strategy;
2) Specify where the revenue and profit come from;
3) Articulate cloud customers' business values, attitudes, and beliefs, which are closely associated with the product or service (such as cloud instance) offerings;
4) Focus on actual business customers' behaviors;
5) Make sense to the firm's senior executive team and the broad;
6) Flexible and quickly accommodate or anticipate changes in markets or consumer behaviors.

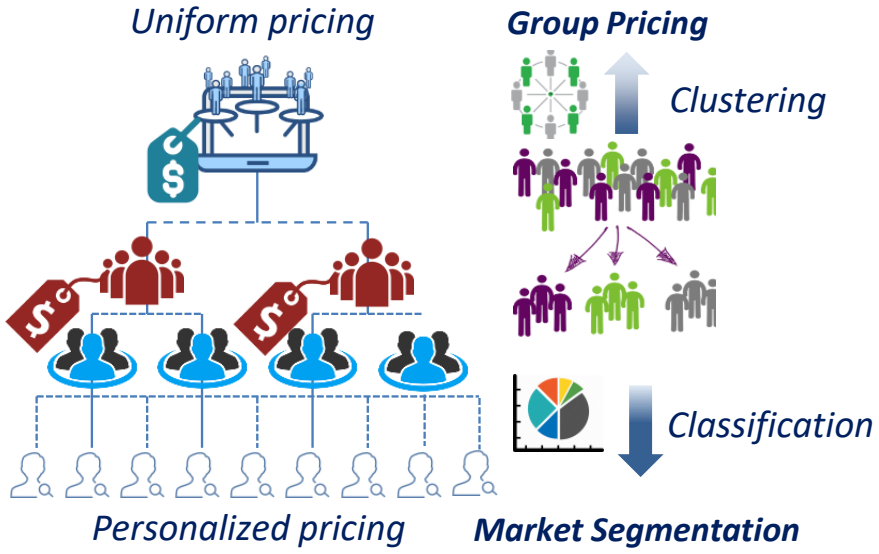


Figure 4—1 Uniform, Group, and Personalized Pricing

Based on these criteria, we develop a novel solution that allows CSPs to identify the B2B cloud market segment quickly. In comparison with other traditional methods, such as analytical[189], strategy-based[190], nested, survey, and Delphi methods [191], it is much more tangible, flexible, agile, and cost-effective for a CSP to roll out different cloud pricing models for its cloud business strategy [187]. It also enables CSP to respond to the ever-changing environment of the cloud market rapidly. The inputs and outputs of the process for our solution are illustrated in Figure 4—2

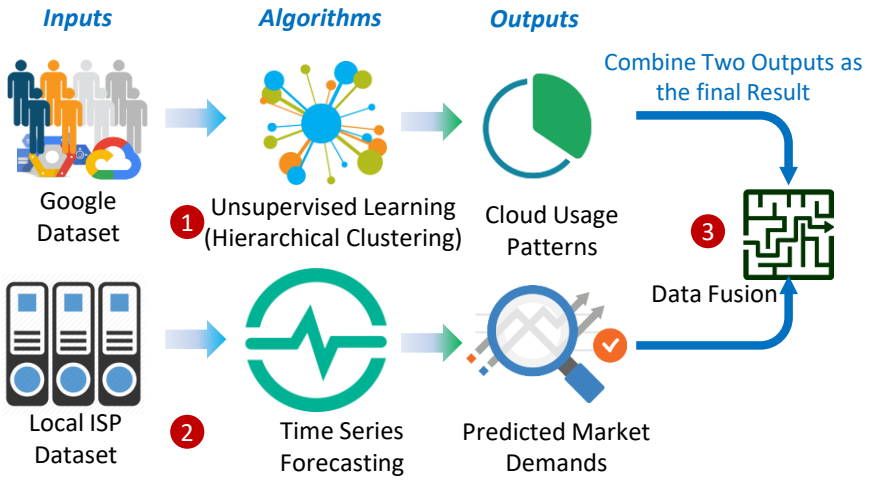


Figure 4—2 the Solution Process of Cloud Market Segmentation

The solution is summarized in three steps:

1) We use hierarchical clustering to segment cloud market;
2) We apply time-series forecasting (TS) for the sales volume prediction;
3) We combine both results for each market segment. We use both Google's and local hosting datasets in our analysis to demonstrate our methodology. The final results are shown in Table 4 — 1.

Table 4—1 The Expected Results of Segmentation

Cloud Market Segmentation	**Segment1**		**Segment k**	**Total**
Demand (Sales Quantity)	q_1	…	q_k	$\sum_{i=1}^{k} q_i = Q_T$
The proportion of Each Segment	$p_1 = q_1/Q_T$	…	$p_k = q_k/Q_T$	$\sum_{i=1}^{k} p_1 = 1$
Market Segment's Charectretsics	Memory Pattern	…	Memory, CPU, Network	k

By doing so, we make three contributions:

1) We demonstrate how to use hierarchical clustering (HC) algorithms to identify the optimal number of cloud market segments.
2) We use TS forecasting to predict the local B2B market demand for virtual machines (VMs).
3) Finally, we combine both results into the final cloud market segmentation table so that a local CSP can leverage it to build different cloud price models for its targeted market.

The rest of the chapter is organized as follows. In Section 4.2, we provide a brief literature review of market segmentation. In Section 4.3, we describe the details of our solution of market segmentation, such as the fundamental principles of the experimental methods and some assumptions that we made. In Section 4.4, we illustrate how to use the HC to segment the cloud market and find the appropriate number of segments. In Section 4.5, we present how to forecast the quantity of VMs demands and then combine both results. In Section 4.6, we analyze and discuss our empirical results. Section 4.7 provides a summary of this chapter.

4.2 Related Work

Since Smith [168] first cast the term of market segment, the topic has been studied in great detail in terms of its theory, methodology [171], concept, foundation, and process [172]. Along with the consumer market, the B2B market theory [173] has also been developed due to its growing momentum and substantial market size and values. Due to the targeted value proposition of the B2B cloud market, namely product, price, place, and promotion (or Kotler's four Ps), the related work consists of theory, analytic approach, and cloud pricing in term of the market segmentation.

According to Wedel [171], the essence of the market segmentation is "a theoretical marketing concept involving artificial groupings of consumers constructed to help managers design and target their strategies." In practice, it is an iterative process to assign a set of variables (e.g., four Ps) to many potential customers that help a firm to form homogenous groups. Under the Wedel's concept, Thomas [202] gave a further clarification of the B2B market segmentation, which is "a dynamic business decision process driven by an (economic) theory of how market functions." In

practice, it is a set of decision processes and activities that can be divided into two different approaches: One is the top-down approach, which is the process of splitting customers into different segments. Another is the bottom-up one, which is to agglomerate each customer into different groups. Claycamp et al. [170] claimed that although the top-down approach is appealing and straightforward, it is very challenging to implement because the splitting process is mainly to drive the potential value of customer surplus. Claycamp exhibited that market segmentation is ultimately a bottom-up process of aggregation in theory. However, the bottom-up approach is also facing challenges in practice because some parameters are very hard to estimate, such as marginal response or managerial requirements [193]. One of the solutions is to propose some controllable marketing variables in identifying marketing stimuli, which is down to only one "P" (Promotion). It is like an analytic approach.

Ralph Oliva [201] indicated the B2B market "segmentation is an analytic discovery process for dividing a large group of customers or prospects into smaller groups." Similarly, Seufert [192] presented an analytic approach to segment user groups for the freemium pricing model. Their approach focused on the core value of the business. If we compare the core value with the hedonic value analysis [138], we can draw an analogy between Irwin Gross' core value, cost, and prices with the hedonic function

$$p_j = \sum_{l=0}^{L} p_l + mc_j + \frac{Q_j}{\left|\partial Q_j / \partial p\right|} \tag{4-1}$$

where p_j is the price of the cloud VM instance "j." mc_j is the marginal cost, Q_j is a quantity, $|\partial Q_j/\partial p|$ is the partial derivative of the quantity taken in term of a price, and $Q_j /|\partial Q_j/\partial p|$ is a markup price, and p_l is the CSP's purchasing price from other vendors. Here, a potential value loss is defined by consumer surplus (CS_i) [7]. The core value of B2B market segmentation is the economic driving force (Figure 4—3)

However, Plank [194] criticized the analytic approaches because many methods are complicated to translate their analytic results into a business strategy. In order to improve the analytic approach, Verhallen et al. [190] proposed a strategy-based approach, which is to identify unobservable characteristics (e.g., firm's goals, objectives, strategy types, and long-term plans) in contrast to observable traits (business size, location, and four Ps). Furthermore, in relation to different input variables, Shapiro et al. [186] proposed the nested- approach, which is to nest from demographics, operating, purchasing, and situational variables to personal characteristics, but the

author indicated their approach could not be generalized. Alternatively, Best [191] offered an expert solution that can become a prior probability of input variables for the market segmentation process.

In contrast to the above methods, Balakrishna's [195] focuses on a solution based on how to better use the industrial market concept for the B2B market segmentation. It is more like a generalized solution for the B2B market. Although these solutions are very persuasive, the main issue remains unsolved, which these solutions are unquantifiable for CSPs to implement their cloud business strategy by rolling out different cloud price models. As a consequence, many recent studies directly focus on cloud pricing models for CSPs to maximize their revenue.

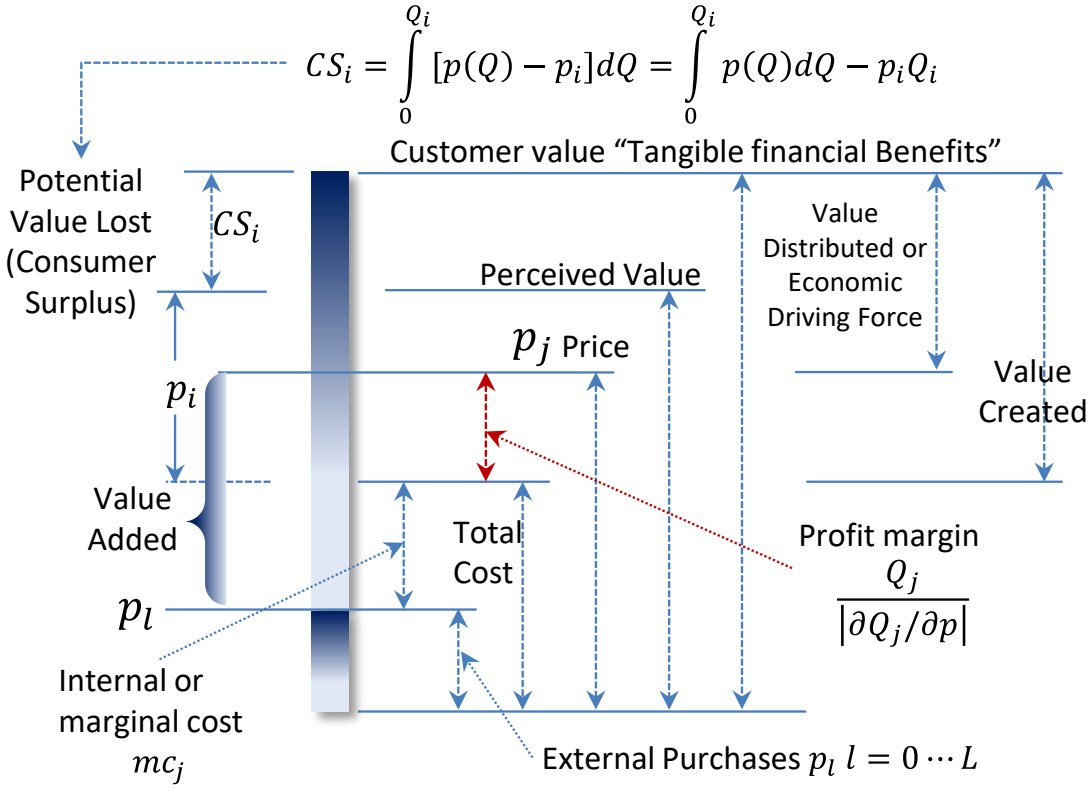


Figure 4—3 Analytic Method of the B2B Market Segmentation

As early as in 2002, Buyya et al. [175] have proposed economic models or pricing schemes to regulate the grid computing resources, which can be considered as one of the prototypes of the cloud pricing model. Javadi et al. [176] developed a statistical model for Amazon Web Service (AWS) spot instance prices in public cloud environments. Although the model is valid, the spot

instance is not desirable for the mainstream of B2B cloud resources because many B2B applications require the mission-critical cloud infrastructure to support its business.

Similarly, Xu et al. [56] proposed the alpha-fair utility function to quantify the applications' needs for cloud users in terms of cloud resource allocation. Although the study is beneficial for a theoretical exploration, the model assumptions require further consolidation because the alpha-fairness utility function is mainly applied to the issue of traffic congestion of communication networks [177] rather than cloud services. Practically, different cloud applications (such as web hosting, database, data storage, virtual desk infrastructure, and so forth) will have different requirements, which lead to different market segments. As respect to the word of segmentation, Wang et al. [165] investigate this problem from an aspect of segmenting cloud capacity, which is to formulate an optimal capacity segmentation strategy for revenue maximization to satisfy the random market demand.

Overall, we can see that there is a gap, which is how to find a quantifiable solution to segment the B2B cloud market so that CSPs can build various optimal price models for their targeted market or customers in connection with both internal costs and external market demand. Our solution provides the answer to this gap.

4.3 Preparation Tests

As Claycamp et al. [170] stated in their theoretical study, the clustering analysis is one of the practical solutions for the market segmentation. However, there are many clustering methods of clustering methods, such as categorical (hard vs. soft), structure (flat vs. hierarchical), data type (model-based vs. cost-based), and regime methods (parametric vs. nonparametric). The question is which one is the right method for our problem.

A good strategy is to explore the datasets in our hands. The first dataset is Google's cloud trace [178], which consists of large cloud clusters for more than 12,500 VMs. It has six dimensions: timestamp, job ID, Task ID, and job type, normalized task cores, and normalized task memory. However, Google has obfuscated some information on the dataset, in which "certain values have been mapped onto a sorted series" for confidential reasons. Fortunately, the encryption schemes will not impact market segmentation because we are looking for underlying customer usage patterns.

The second dataset is collected by one of the leading Australian telco firms for its hosting business. The dataset has sales records of web servers for its business customers between 2003 and 2009. The idea of the first experiment is to estimate the number of cloud market segments and the proportion of each segment. Google's dataset would unveil cloud usage patterns. We assume that both global and local cloud customers have the same usage pattern in this case. The 2nd experiment is to forecast the local B2B market demand because the B2B market demand is closely associated with a robust local B2B relationship [179].

4.3.1 **Proposed Method of Segmenting**

On the base of the good criteria for the segmenting market [169] and the dimensions of Google's dataset, we propose the HC method. The reasons are as follows:

1) We do not know the exact number of cloud market segments in advance.
2) Referring to Claycamp's theory [170], it has to be an agglomerative process of fusion clustering, which is a bottom-up process of clustering.
3) Furthermore, it would be preferable to leverage HC because we can form a dendrogram (tree diagram) that allows us to choose the dendrogram at any desired level. These analytics features allow CSPs to segment the B2B market at any granularity level so that a CSP can explore opportunities of any niche market.

However, all methods have their disadvantages. One of the primary difficulties of HC is too sensitive to the number of clusters. One solution to solve this problem is to use Ward's algorithm to minimize the variance of Sum Square of Errors (SSE) by consideration of all possible methods. Our overall strategy of the 1st experiment is illustrated in Figure 4—4. The essence of the clustering algorithms is to calculate dissimilarity that is measured by the Euclidean distance of data points. For the Ward's algorithm, the equations of SSE are as follows:

$$\Delta_{C_a \cup C_b} = SSE_{C_a \cup C_b} - \left(SSE_{C_a} + SSE_{C_b}\right) = \frac{n_a n_b}{n_a + n_b}(\mu_a + \mu_b)^2 \tag{4-2}$$

$$where\ SSE_{C_a} = \sum_{i=1}^{n_a}(a_i + \mu_a)^2\ ,\ SSE_{C_b} = \sum_{i=1}^{n_b}(b_i + \mu_b)^2\ , nd\ SSE_{C_a \cup C_b} = \sum_{i=1}^{n_c = n_a + n_b}(c_i + \mu_c)^2 \tag{4-3}$$

where $\Delta_{Ca \cup Cb}$ is the cost function to combine two clusters C_a and C_b that have the number of observations n_a and n_b, respectively. a_i, b_i, and c_i are the ith observations in the cluster C_a and C_b,

and the merged cluster $C_a \cup C_b$. Likewise, μ_a, μ_b, and μ_c are the centroid of these clusters. To update the Euclidean distance in Figure 4—4, we can use the Lance-Williams dissimilarity update formula [198].

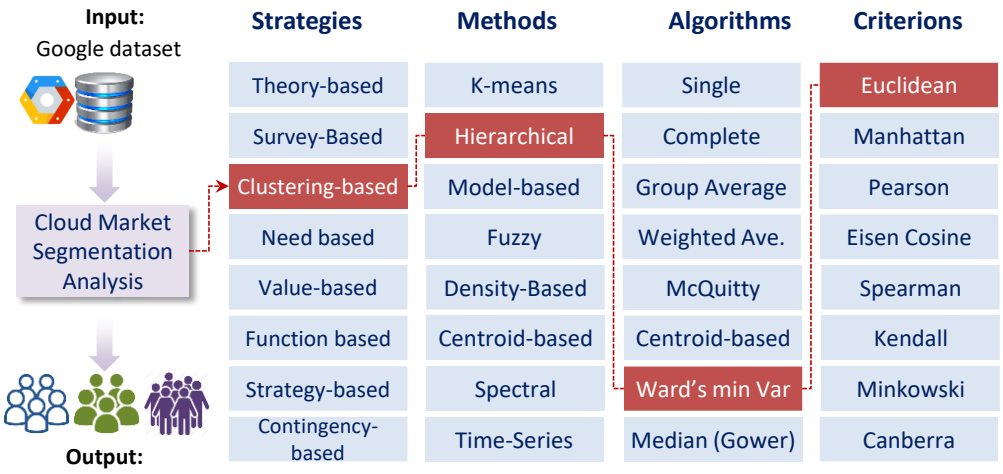


Figure 4—4 The Map of Hierarchical Clustering Method

4.3.2 Proposed Method of Prediction

The idea of the second test is to predict or forecast the B2B market demand in the next 12 months so that we can build cloud infrastructure capacity to meet the local cloud market demand. Several techniques can be applied for prediction, such as linear and multiple regression, random forest, decision tree, ANN, and time series forecast.

In this study, we adopt the time series forecast model to predict the total volume of VM sales. The reasons are:

1. Time series forecasting is simple. It would be easier to be presented to the firm's executive team.
2. We can estimate each sales volume for every month or year so that it would be convenient for cloud capacity planning.
3. The forecasting result will tell the confidence interval.
4. It can be updated very quickly.

4.4 Cloud Market Segments

We test Google's dataset first and see whether the dataset has meaningful patterns or not. This process is called the "clustering tendency evaluation." The reason to check the clustering tendency of the data is that a hierarchical clustering method can impose patterns or clusters onto a randomly distributed dataset even if there are no such definable or extractable clusters within the dataset.

Liang and Kotagiri et al. [180], [181] did some studies regarding clustering tendency assessment. There are many techniques available for cluster tendency evaluation. One of the methods is Hopkins statistic [182] null hypothesis test. Hopkins' test can be expressed using the following equation:

$$H = \frac{\sum_{i=1}^{n} P_i^2}{\sum_{i=1}^{n} I_i^2 + \sum_{i=1}^{n} P_i^2} \tag{4-4}$$

where I_i square is the distance between an observation x_i and its nearest neighbor x_j (x_i, $x_j \in D$ = dataset). P_i square is the distance between a random y_i and its nearest neighbor y_j (y_i, $y_j \in D_r$ = random dataset). The null hypothesis test shows that if H value is equal or close to 0.5, the tested dataset D has no meaningful clusters so that we accept the null hypothesis. Otherwise, we reject the null hypothesis. Based on the above Hopkins' equation, we calculate the Hopkins' index value. It is equal to 0.064, which is approaching zero. We also use R command "fviz_dist" (display dissimilarity matrix) to visualize Google's dataset with a comparison of a randomly generated dataset (Figure. 4-5 on the left)

Figure 4—5 Assessing Clustering Tendency of Google's Dataset

The pink color indicates I_i square = 0, and the purple color means I_i square = 1. In contrast, the right diagram of Figure 4—5 shows that both values are randomly distributed

across the dissimilarity matrix. Hopkins null hypothesis test result tells us Google's dataset has a clustering tendency.

4.4.1 **Extract Cloud Usage Patterns**

For the R system, the bottom-up and top-down are known as Agglomerative Nesting (or AGNES) and Divisive Analysis (or DIANA), respectively. The linkage algorithm is "Ward' because we want to minimize the SSE variance. If we temporarily assume the number of segments is four (McDonald [172] suggested the number is between 5-10 and others suggestion is between 4 and 5 [183]), we can plot out the dendrogram or segment (Figure 4—6).

We can also cut the cluster dendrogram into seven segments by moving the vertical distance height around to height distance 10. Consequently, clusters 1 and 4 are split further, and 2 and 3 remain the same (Figure.4-6). The number of clusters seems to be decided arbitrarily. Now, the issue is how we chose an optimal number of clusters, "k."

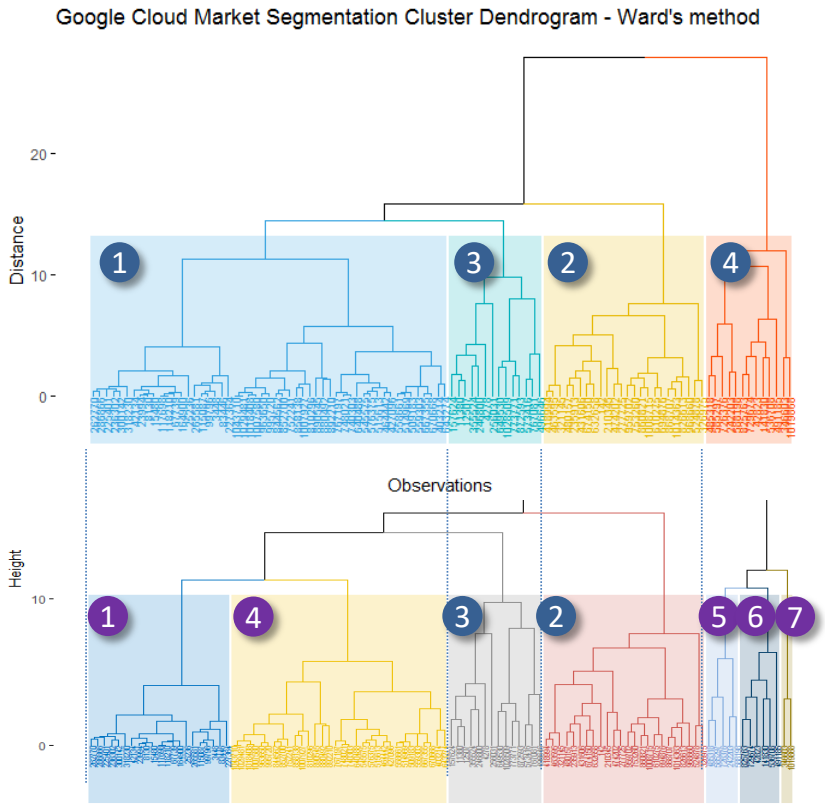


Figure 4—6 the Result of Cloud Market Segmentation

4.4.2 **Deciding the Optimal Number**

This is a challenging question. If the number is predetermined, we can adopt other algorithms to do the clustering, such as k-means. However, this number is unknown. Fortunately, many existing schemes can help us to estimate this number, such as Dark Block Extraction (DBE) [184], hierarchical, partitioning, direct, statistical testing, density mode seeking, clumping, grid-based clustering, etc. R has more than 30 methods or indices to decide this optimal number. Charrad et al. [185] developed the "NbClust" package to decide the number of clustering. Our analysis of Google data shows the optimal number "k" is four (Figure 4—7).

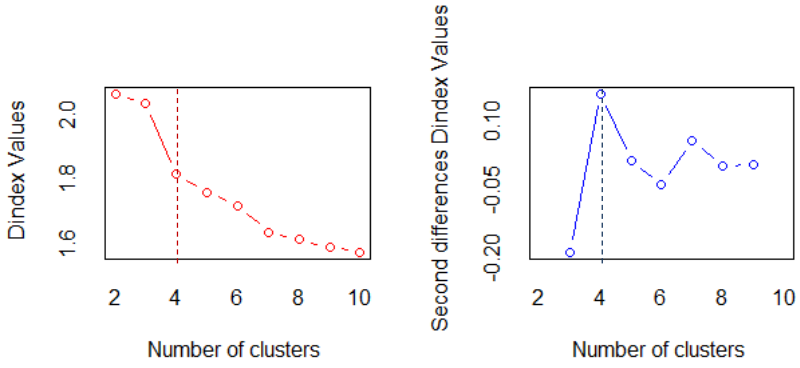


Figure 4—7 Optimal Number of Test Result by NbClust Package

The index is shown in Figure 4—7 is the Dindex graphic to determine the optimal number of clusters. Dindex is to measure clustering gain on intra-cluster inertia [185], which is the degree of homogeneity between the data points in a cluster. The equation of Dindex can be presented as follows:

$$w(P^q) = \frac{1}{q}\sum_{k=1}^{q}\frac{1}{n_k}\sum_{x_i \in C_k} d(x_i, c_k) \tag{4-5}$$

$$gain = w(P^{q-1}) - w(P^q) \tag{4-6}$$

where P^q is the "q" number of partitions by imposing "k" number of clusters, "d" is the distance and "c_k" is the center of a cluster, "n_k" is the number of data points in a cluster. "x_i" is any data point within a cluster. The clustering gain on intra-cluster inertia should be minimized. Ultimately, the Dindex is to measure "the degree of homogeneity of the data in a cluster." [185]

4.5 Demand Prediction

Oliva [201] indicated any B2B market strategy has to focus on the object of the Key Account Market (KAM). In this case, ISP has to predict its own local cloud market demand so that the ISP can achieve a realistic sales forecast. This target can be either arbitrarily or rational. If an executive team requires a making- sense sales target, the forecast demand has to come from a local dataset.

For the local ISP firm, the natural extension of the cloud business is its existing web hosting business. It can leverage its previous sales records to estimate the cloud market demand. Our second dataset has 3,192 data points (Windows servers only) over 67 months (between Aug-2003 and Feb-2009). We can plot the hosting server sales volume monthly (Figure.4-8).

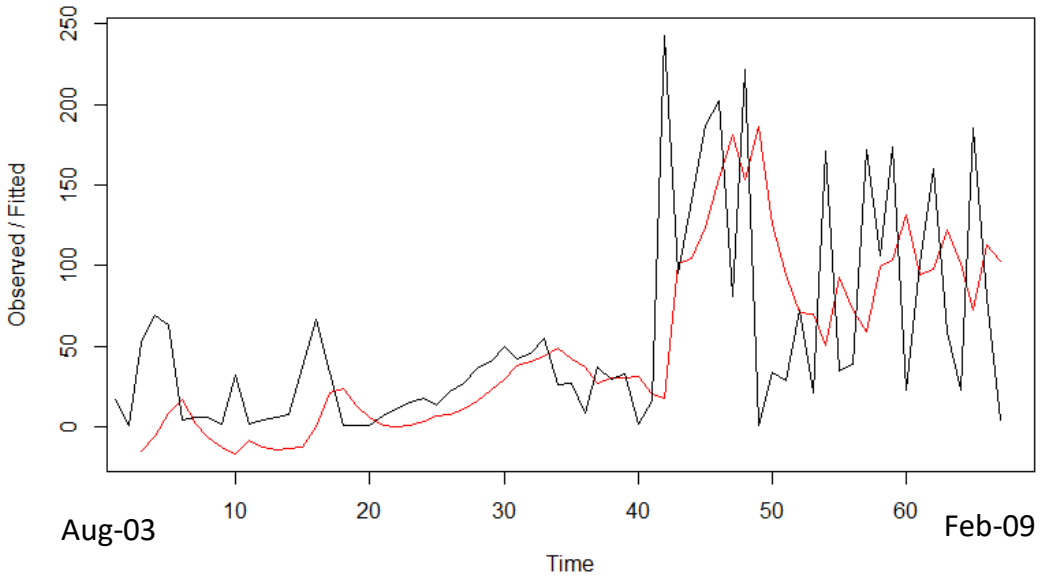


Figure 4—8 Local Hosting Service Monthly Dataset

The red line in Figure 4—8 is to smooth the observation data points. As we can see it, the sales volume is quite low in the first 40 months but the movement of the next 27 months was very volatile.

There are many different methods to estimate or predict the future sales volume, such as logistic regression, support vector machine (SVM), decision trees or Classification and Regression Tree (CART), random forests, and time series (TS). In comparison, TS [187] would be a better tool to estimate the sales volume because the dataset is collected in a time series. Moreover, it can give

us the monthly and yearly forecasting quantity or VM sales. This is what our goal of this 2nd test. It will also be valuable for cloud capacity planning and budgeting.

Although the seasonal component is not apparent, we still set the "gamma" value equals to "False" to remove the seasonal components in the TS model. We then use the "forecast" package of R to plot the next 12 months (Figure 4—9, left) and eight years of trends (Figure.4-9, right). We can see there is a downward trend in sales volume for the monthly but upward trend for the yearly forecasts.

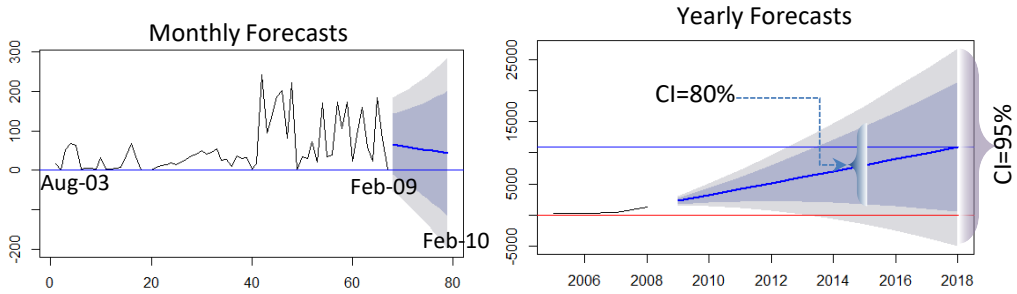


Figure 4—9 VM Sales Prediction Results

Now, the issue "Is the TS a valid model for the forecasting?" We can plot the model residual to visualize the error trend. If we find any pattern in the residual plot, it means the model is inadequate for prediction. Otherwise, it is a good TS model. Based on Figure 4—10, we can see the residuals are moving around zero.

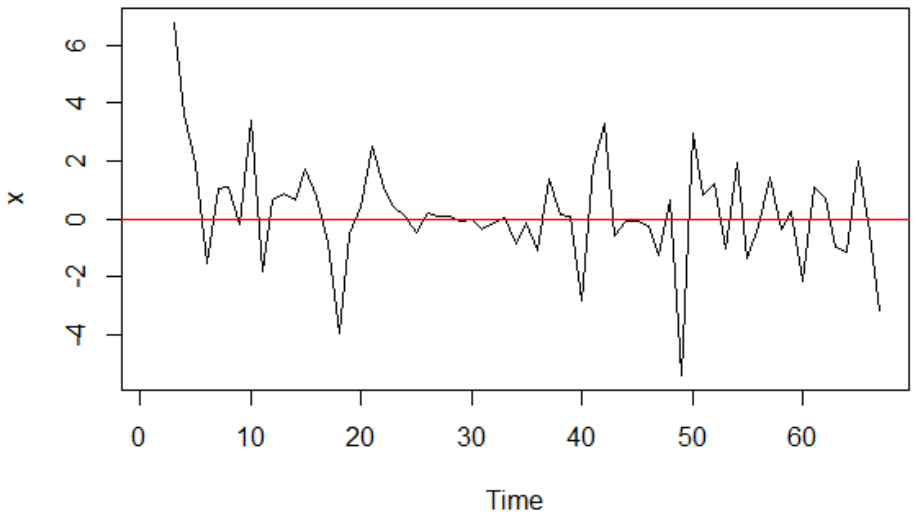


Figure 4—10 Residuals of TS model Sales Volume

To validate this TS model, we can use both the histogram plot and Auto-Correction Function (ACF) function (Figure 4—11). The histogram plot (left of Figure 4—11) shows a normal distribution and the ACF plot (right of Figure 4—11) shows there is only one line that exceeds the boundary limit lines. So, we can conclude the TS model is valid

If we adopt recent Gartner's reports to assume the average market share of Windows server is around 36.56%, we can estimate the final result of total VM quantity is 6,250 in 2009 (2,285 for Windows servers) as noted in Table 4—2

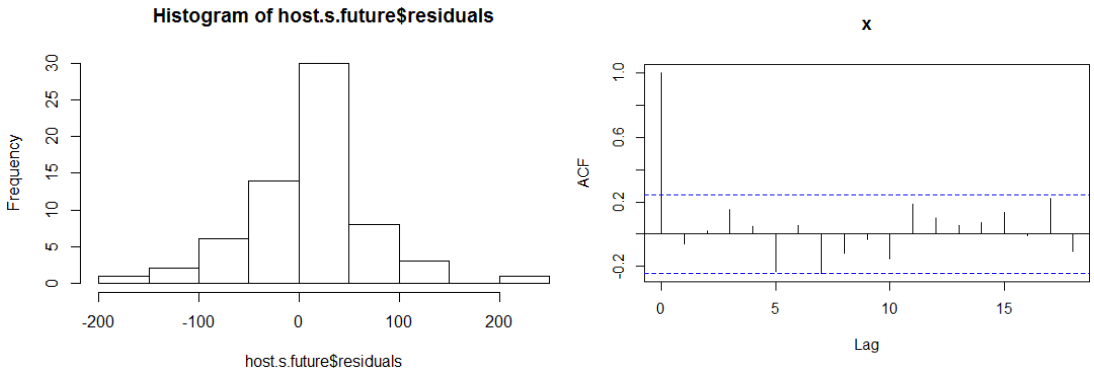


Figure 4—11 TS Residuals Histogram and ACF plot

Table 4—2 Yearly Forecasts VM SALEs

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Win. Servers	2,285	3,241	4,197	5,153	6,109	7,065	8,021	8,977	9,933	10,889
All VMs Qty.	6,250	8,865	11,480	14,095	16,710	19,324	21,939	24,554	27169	29,784

As per the solution noted in Table 4—1, we combine two test results for the final market segmentation are shown in Table 4—3

Table 4—3 Final Result of Market Segmentation

Segment	Seg. 1	Seg. 2	Seg. 3	Seg. 4
Job Priority	2	1	0	3
Cores	1	1	23	11
Memory	6	5	6	99
%	10.05%	56.46%	22.97%	10.53%
Sales Vol.	593	3329	1354	620
Possible Workload	Static	Dynamic	High Availability	Backend

4.6 Analysis and Discussion

Our three-step process solution shows how to segment the B2B cloud market for the ISP to expand its existing business from hosting to the cloud. The novelty of our solution is that it can practically extract the cloud customer usage patterns from Google's dataset. The job priority (as shown in Table 4—3) means the scheduling constraints on some jobs. The most substantial proportion of cloud usage (or workload) is segment 2, in which most customers were using only one core and a lower amount of memory. It is not surprising that Google indicates users often overestimate their resource consumption. In contrast, the lower priority jobs (or backend data processing) consume the most significant amount of memory capacity (Segment 4). Although the top priority job of segment 3 consumes a lot of computing power (23 cores), memory usage (6) is relatively less.

Based on the limited parameters shown above Table 4—3, we can probably guess what type of workload is most likely even though Google data did not provide this information. Segment 1 is more like static web hosting workload; Segment 2 would be dynamic (because of job priority ranking is high than static). Segment 3 is more like a Highly Availability workload, such as customer relationship management (CRM) applications, and segment 4 is more like backend workloads, such as database backup or business analytics. One of the insights from Table 4—3 is the cloud infrastructure, or a server farm should be tailored into 12 units per cloud server cluster. A memory configuration should be built in 6 GB per slot.

For the HC algorithm, it is essential to indicate that one of the influencing factors for the optimal number of the market segment is "seed," However, it does not only impact on the clustering method but also other methods that require setting "seed." In this study, we assume there are no differences regarding usage patterns between B2C and B2B for Google's dataset. By using the HC algorithm, we can meet the good market segment criteria [169] 3, 4 and 6. However, the HC algorithm alone is not enough because the input dataset comes from global CSP. It only provides the cloud customer behaviors.

The total cloud market demand estimation has to come from a local B2B dataset. Typically, the sales target often becomes the Key Performance Index (KPI) for a senior management team. It is desirable to use a TS model for the local market demand because the B2B cloud market is often built upon the long-term B2B relationship. Furthermore, the purchasing decision is made by a

group of people rather than a single individual. The TS can deliver both monthly and yearly sales forecasts. By adopting the TS model, we can satisfy the criteria [169] of proper market segmentation 1, 2, and 5. In comparison with other solutions (Table 4—4), this solution has the following advantages:

Table 4—4 Segmentation Solution Comparison

Different Methods for Market Segmentation	Customers' Business Values	Focus Usage pattern	Flexible	Align with business strategy	Specify revenue and profit	Make sense
Analytic Method	√		√		√	√
Nested Method	√	√				√
Strategy-Based			√	√		
Delphi Method		√		√		√
HC + TS	√	√	√	√	√	√

- The solution is practicable and quantifiable, which has the input variables (Table 4—4) for the process of the B2B cloud market segmentation.
- The solution can quickly be updated for the rapidly changing environment of the cloud market, such as customer behaviors shift, the internal investment budge variation, and the cloud technology eruption.
- It can assist senior executives in a managerial decision to test different local niche markets that many global CSPs might not have a local B2B relationship.
- The solution allows CSP to develop a pricing model based on both the market and customer-value, which emphasizes both the external rationality rather than internal rationality.

In contrast, the analytic method cannot extract usage patterns, and the nested approach has to be case-by-case. The strategy-based method is often quite challenging to be translated into a practical solution. Survey and Delphi methods often take too long to be accomplished and often, it is indirect.

To the best of our knowledge, it is the first kind of study on the B2B cloud market segment. Many existing and incoming CSPs require this kind of knowledge to assist their cloud business investment strategy in terms of budgeting and resource capacity planning. Market segmentation helps CSPs to find a better pricing strategy for maximizing their profits.

4.7 Summary

This chapter shows how to segment the cloud market in three steps. In comparison with other the market segmentation approaches such as nested, analytic, Delphi, and strategy-based approaches, this method is tangible, quantifiable, and compelling, just as P. W. Bridgman philosophically argued if we don't know how to measure it, we really don't know what it means[203].

Overall, this chapter proposes a novel solution that combines both hierarchical clustering and time series forecasting on the basis of the classical theory of market segmentation. The tested results and empirical analysis show that this solution can efficiently segment cloud markets and also predict the market demands. It lays out the groundwork for value-based pricing of baseline cloud services

Chapter 5

Modeling Cloud Customers' Utility Functions

Modeling cloud business customers' utilities is one of the critical issues faced by many cloud service providers (CSPs). It concerns how to measure various subjective experiences of the business customers and how to translate their cloud service experiences into a quantifiable unit, which can be determined by a specified utility function for cloud resource consumption. The aim of this quantification is to set up a pricing foundation so that CSPs can capture a broader range of utilities from different market segments and identify the optimal price point of each pricing model to maximize the cloud business profits for its pricing strategy. Previous studies either focused on simple theoretical proof or drifted the meaning of utility between demand and supply or proposed a solution based on a single cloud market. This chapter proposes a novel and practical solution to model multiple utility functions for various business applications based on a scenario of six cloud market segments, which are analyzed by three analytic approaches, namely Highly Availability (HA) analyzed by Markov chains, online e-commerce analyzed by queueing theory, and backup and backend analyzed by risk assessment. This modeling method emphasizes the value of co-creation with cloud business customers. In comparison with other methods, such as calibrated, price-quality, resource-based, simple linear, and capacity-aware, this method provides both internal and external rationalities for CSP's pricing strategy to gain more than 83% of cloud market shares while other methods can only achieve less than 17%.

5.1 Introduction

The goal of this research is to define various utility functions for different cloud business customers within different market segments so that a Cloud Service Provider (CSP) can capture a broad spectrum of cloud market share and revenues to maximize its profit. Moreover, the CSP can tailor its limited resources to serve its target customers more effectively.

This Chapter is derived from:

- **Caesar Wu**, Rajkumar Buyya, and Kotagiri Ramamohanarao, "Modeling Cloud Customers' Utility Functions," *Journal of Future Generation Computer Systems(FGCS),* Volume 105, Pages: 737-753, ISSN: 0167-739X, Elsevier Press, Amsterdam, The Netherlands, April 2020.
- **Caesar Wu**, Rajkumar Buyya and Ramamohanarao Kotagiri, Big Data Analytics = Machine Learning + Cloud Computing, Big Data: Principles and Paradigms, ISBN: 9780128053942, Waltham, MA Morgan Kaufmann, Elsevier, 2016. p.3-37

The word “utility” is very ambiguous and often confusing. One of the primary reasons is it has many connotations[218]. The common sense of utility means “the usefulness of something, especially in a practical way.” For example, the utility of database means to implement various processes or functions of the database, such as batch update, rebuild, recovery, backup, etc. Another sense of utility is quite close to the meaning of the usefulness that often refers to the basic infrastructure of public services that are offered by incumbent service providers, which is called “public utility” or simply, “utility.” Buyya et al., [52] defined the infrastructure of “cloud computing” as the 5th utility. Still, another meaning is the utilization rate, which means the effective usage of something.

Our definition of utility is in an economic sense, which is to focus on a particular consequence of an individual’s decision making. This consequence is measured by the individual’s subjective satisfaction, happiness, and worthiness for the goods and services to be consumed. These subjective measurements of the utility reflect on a price that the individual is willing to pay for [225]. The acceptable price leads to an idea of modeling various utility functions. It defines a relationship between a price to be paid and a number of goods and services (such as Virtual Machine or VM) to be acquired. According to Krugman and Wells [224], the different individuals would have different utility functions because different people would have different needs and preferences towards a certain amount of goods or services.

The essence of a utility function is to describe how people consume various amounts of goods and services in term of their subjective preference, needs, and experiences in a less or more rational way that is measured by either cardinal or ordinal approaches (“cardinal” measurement means the utility value can be quantified by a marginal value (e.g., an additional subjective satisfaction for one more unit of cloud resources is acquired) and “ordinal” method can only be measured by a ranking or ordering approach). To the cloud computing services, we adopt a cardinal approach [218] to quantify the cloud utility values because the cloud utility measurement satisfies the criteria of cardinal analysis:

1) The cloud business customers are rational,
2) Utility value can be measured numerically in term of dollar value,
3) The unit of Infrastructure as a Service (IaaS) is homogeneity.

So, the focal point of this study is to model different types of utilities (value functions) for various cloud business applications that are classified into different market segments. In particular,

we target the applications of web hosting, content delivery, e-commerce (online check out system), database backup, disaster recovery (DR), virtual desktop infrastructure (VDI), backend, etc.

If we assume that the measurement of the customer's satisfaction (this metric is directly related to the business customer's revenues in term of utilizing or running the business applications), then our modeling process is to estimate how much the customers are willing to pay for a given quantity of the cloud resources so that the customers can improve its business revenue. Figure.5-1 highlights the entire framework of a value-based cloud pricing strategy and how the 2nd step of the modeling utility function is fit into a big picture to achieve the goal of value co-creation with cloud business customers [238].

According to T. Nagle et al. [10], this is a challenging task because the issue requires multidisciplinary knowledge. Many previous works [206] [207] [208] [209] [210] [213] [214] [215] have made a lot of excellent progress for this problem. However, there is still a significant gap in how to apply some previous modeling methods in practice. The gap is often caused by an ambiguous definition of utility in the first place. Some modeling methods often mix a demand side's utility value with a supplier side's price or cost. Most importantly, utility models assume a single market only without consideration of market segmentation.

To overcome these problems, this study will emphasize on the demand side of utility values and solve the problem by carving out the big problem of pricing into four smaller and manageable issues: cloud market segmentation (See Chapter 4), utility functions modeling for cloud business customers, cloud pricing modeling and cloud price optimization for CSP to achieve maximum profits (Figure 5—1). This chapter only deals with the issue of modeling multiple utility functions (Step 2) for cloud business customers (not for CSPs or end-users). The issue of cloud market segmentation (Step 1) has been discussed in early work in chapter 4. The other two issues (Step 3 and 4) will be addressed separately in the following chapter.

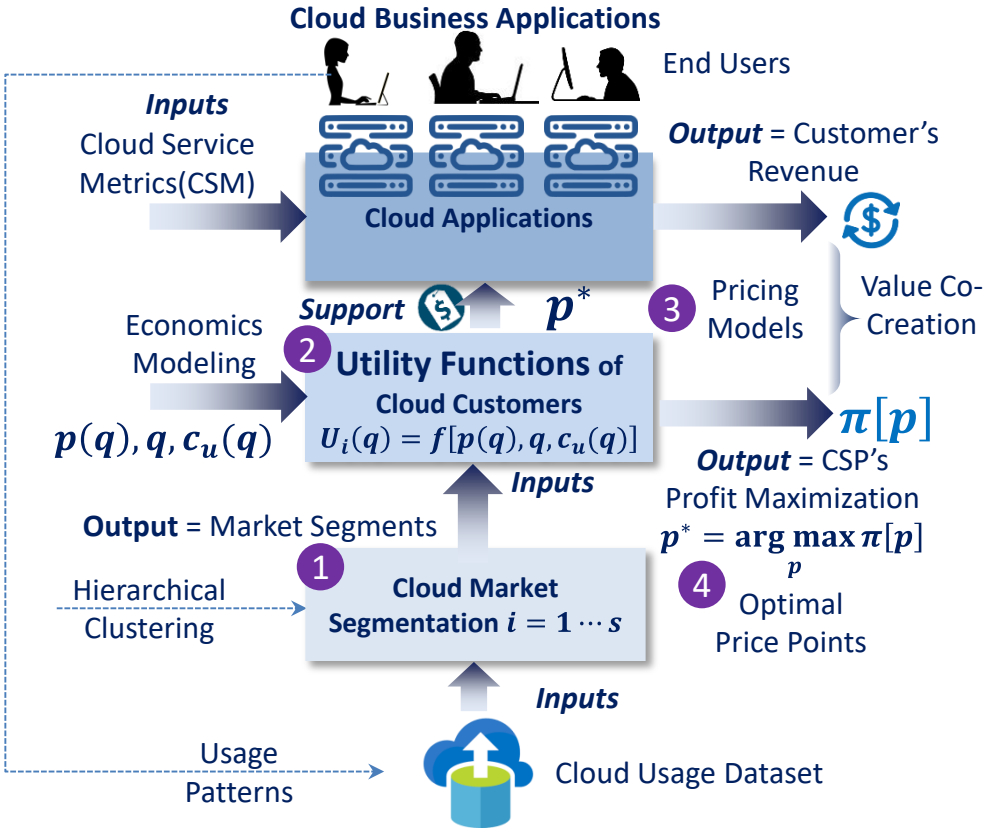


Figure 5—1 Model Cloud Utility Functions and its Measurement

According to the economic definition of utility, we can clarify the meaning of the utility for cloud services, which is a subjective measurement of the cloud customers' values or satisfaction for the number of VMs to be consumed. We can also use the cloud customer's experiences (CX) or key performance indicators (KPI) or cloud service metrics (CSM) to measure a customer's subjective values. Both the National Institute of Standards and Technology (NIST) [205] and Oracle [227] have defined CX, KPI, and CSM along with three tangible business dimensions that consist of acquisition (increase in sales), retention (monetize relationships), and efficiency (leverage investments). All of the quantitative measurements for business dimensions can be translated into the cloud business customers' revenue and profit improvement. To articulate our modeling process clearly, we consider a real scenario of how a CSP's to develop its cloud pricing strategy in terms of developing its cloud business plan.

5.1.1 Motivation Scenario

Suppose a board of directors of a hosting firm (supply side) decides to expand its traditional hosting market to the local cloud Business to Business (B2B or industrial) market (demand side)

for the goal of increasing both revenues and profit with a fixed amount investment budget. If the firm understands its own technical expertise (capability) well and identifies its targeted customers clearly, the subsequent issue is how to segment the B2B market for its addressable or potential market.

5.1.2 **Problem Definition and Solution**

The purpose of identifying the market segment is to find a solution on how to serve the targeted customers well for a limited resource or investment budget so that the firm can achieve sustainable business growth. Ideally, a CSP should make every customer pays a different price so that it can extract the maximum utility value from each customer [170]. This pricing strategy is the so-called price of perfect discrimination. However, it would be too costly to do so. The alternative way is to group targeted customers who have the same characteristics together. This idea leads to "market segmentation." If we assume that the firm has completed the process of market segmentation and identified six segments as shown in Table 5—1 by leveraging Google's public dataset [178], we can probably find there are six possible cloud market segments based on the various parameters or characteristics of cloud customers' usage patterns. Figure. 5—1 shows the result of the market cloud segmentation (cluster dendrogram). The process of how to identify these market segments can be found in Chapter 4. The decision to adopt the scenario of six market segments can be justified as follows:

Table 5—1 Defining Cloud Customer Utility Functions

Segment	**Seg 1 $U_1(q)$**	**Seg 2 $U_2(q)$**	**Seg 3 $U_3(q)$**	**Seg 4 $U_4(q)$**	**Seg 5 $U_5(q)$**	**Seg 6$U_6(q)$**	**Total**
Average Job Priority	1	0	2	0	3	3	
Average number of Cores	2	23	1	1	13	3	
Average number of Memory	7	6	6	3	102	86	
Perentage	30.1%	23.0%	10.0%	26.3%	9.1%	1.4%	100%
Predicted Sales Vol	269	205	90	235	81	13	893
Estimated Possible Workload	Static or Dynamic	Static or Dynamic	Static	HA	HA	Backend	
Example of Apps	Web Hosting Server & Online checkout	Dynamic Content Delivery	Virtualized Desktop Infrastructure	Database Backup Server	Disaster Recovery & BI	Logfile process	

In order to model a utility function, we begin to ask how to identify the optimal price point for CSP's achieve profit maximization from a perspective of value co-creation. Figure 5—2 provides details of a processing solution for modeling multiple utility functions by various analytic approaches and business requirements for different business applications. In comparison with other modeling methods, such as empirically calibrated [251], price and capacity [231], resource optimization [209], response time, capacity-aware [229], utility-based-SLA [69], and Model-Based [206], our modeling method has a number of advantages:

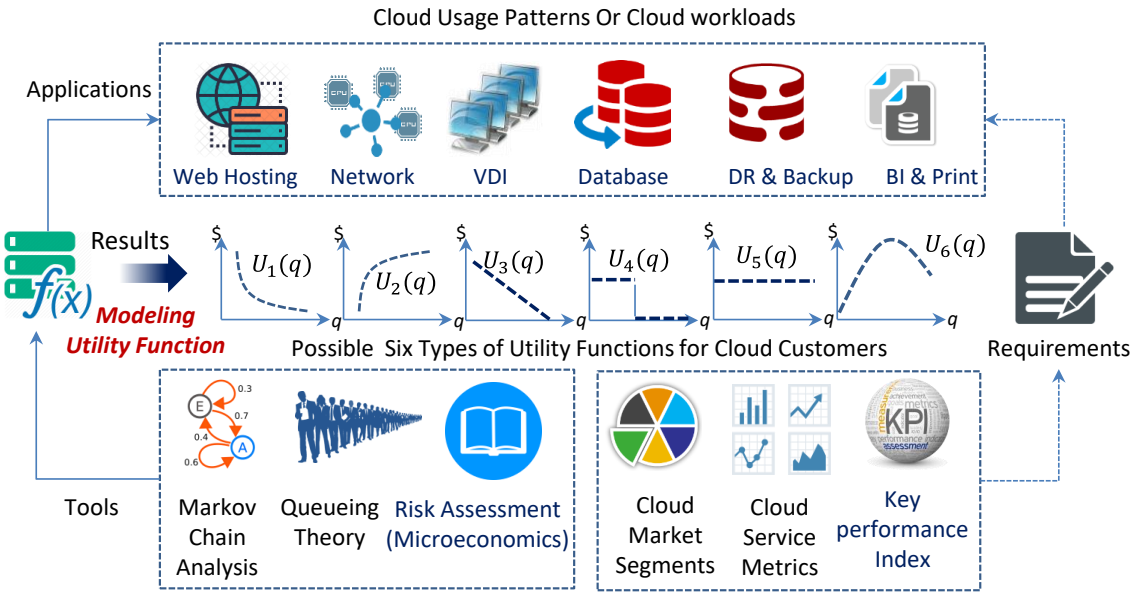


Figure 5—2 The approach to Modeling Cloud Customer Utility Functions

1. It is practical and quantifiable for real business applications,
2. It can be implemented by any CSP for its targeted market,
3. It is derived from the foundation of economics
4. It is agile and flexible to cope with a CSP's business strategy and market segment changes,
5. These utility functions are defined for improving the cloud business customer's revenue and profit.
6. It is a process of value co-creation for both CSP and cloud customers.
7. It gains more market share for CSPs to achieve more profits by optimizing different cloud price models.

We argue our solution is easy and practical for decision-makers to make a critical investment decision in terms of the cloud business. With the listed advantages of our modeling method, we make the following contributions to the cloud paradigm.

5.1.3 **Our Main Contributions of This Work**

To the best of our knowledge, this is the first such study to propose a solution of multiple utility functions under one framework of a segmented cloud market along with different cloud applications. It does not only focus on the modeling utility value of cloud business customers (demand side) but also emphasizes on value co-creation.

Building on the previous result of cloud market segmentation, our novel solution enables CSPs to capture more market share than a single market solution. Consequently, it can deliver a much higher profit margin for CSPs.

Although there are many different units of subjective measurements in terms of CX, KPI, and CSM, this solution is the first time to unify various subjective measurements into a single measurable unit – a dollar that represents customers' revenue and profit improvement. It is a direct, tangible and practicable for cloud practitioners. This modeling solution allows CSP to build multiple utility functions in a single dependent variable with a single independent variable.

Most importantly, this study lays out one of two cornerstones for CSP to define a better pricing strategy from a customer's value proposition. It means that multiple utility functions can be quantified and validated by both internal and external rationalities for CSP to achieve profit maximization.

The rest of the chapter is organized as follows: Section 2 presents how we model multiple utility functions based on the previous result of six market segments and how we make the assumptions and determine the scaling coefficient and other parameters for various utility functions. Section 3 gives a brief review of related works regarding modeling methods for the utility functions. Section 4 provides a detailed performance evaluation and validation for our modeling solution. Section 5 offers the number of guidelines for how to select these utility functions. Section 6 makes our conclusions and outlines future work.

5.2 **Modeling Utility Functions**

As Figure. 5-1 has illustrated, cloud market segmentation is another cornerstone for CSP to build its pricing strategy. In section 5.1.1, we have presented Table 5—1 for six market segments. The issue is how the cloud market segment has been defined? How are six cloud market segments determined? These are critical assumptions of modeling multiple utility functions that have to be clarified first.

5.2.1 **Key Assumptions of Cloud Market Segments**

If we apply a hierarchical clustering method to extract usage patterns from Google's public dataset [178], we can identify there are six possible cloud market segments based on the various parameters or characteristics of cloud customers' usage patterns [232][233]. Figure 5—3 shows the result of the cloud market segmentation (a clustering dendrogram). The process of how to identify these market segments has been discussed in our early chapter. The decision of six market segments scenario can be justified by the following criteria:

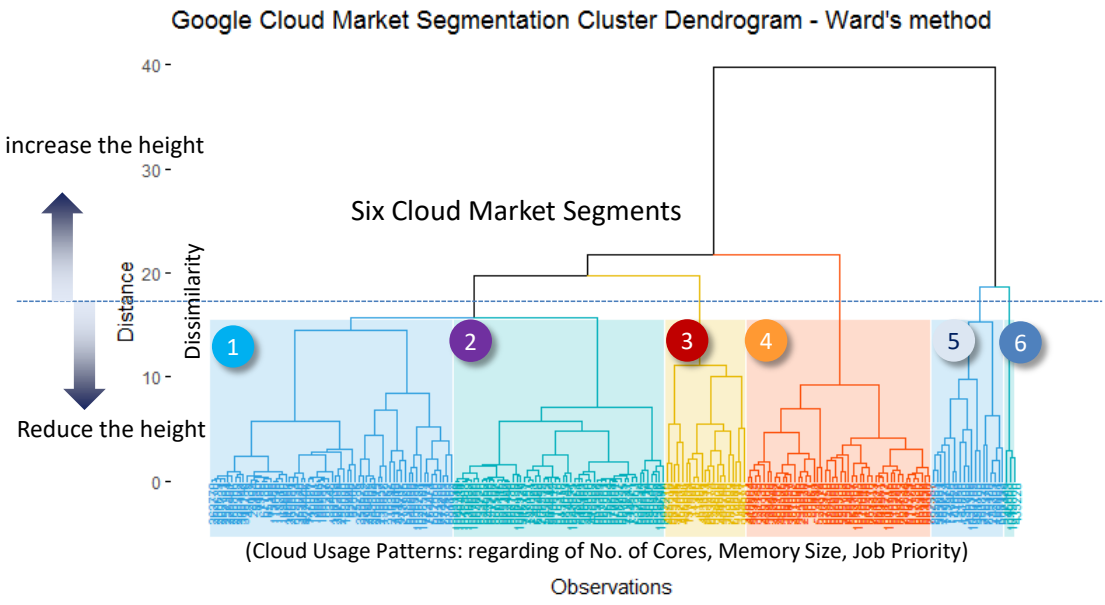


Figure 5—3 Proposed Six Cloud Market Segments

1.) The optimal number of segments is between 4 and 8 by a hierarchical clustering process in Chapter 4.
2.) McDonald [172] suggested that the number of the market segment should be between 5 and 10.

3.) We assume the firm is a traditional hosting company that wants to explore the cloud market, which means the firm is a newcomer to a cloud. The investment budget is limited.
4.) The cloud business strategy is to avoid some higher risky niche market segments.

If the firm's cloud business strategy wants to meet all the above criteria, the number of market segments has to be six (Shown in Figure 5—3) because the above criteria 1 gave the value from 4 to 8 and the criteria 2 suggested between 5 and 10, while the criteria 3 and 4 limited the value at the lower end (5 and 6). From Figure 5—3, the value of clusters is an even number by reducing the height of the cluster dendrogram. Therefore, we adopt six market segments in this scenario.

If a firm is a current CSP that has more investment budget and attempts to explore more risky niche cloud market segments, it can determine to have more than six market segments. The bottom line is that the CSP should clarify its cloud business strategy and targeted customers first, and then the optimal number of cloud market segments can be determined.

5.2.2 **Assumptions of Business Applications**

Furthermore, we can also assume cloud customers' resource consumption (e.g., a particular configuration of VM, workload priority, the number of cores and memory size) is closely associated with a particular business application (e.g., web hosting, e-commerce, database backup, backend processing, content delivery, etc.). Consequently, we approximately establish a corresponding relation between each cloud market segment and a particular cloud workload pattern (See Chapter 5).

The assumption of the six market segments only gives one type of cloud business scenario. If a firm is one of the existing CSPs and attempts to explore more new niche market segments and has more investment budget, it can decide to have more market segments. The bottom line is that the CSP should clarify its cloud business strategy and targeted customers first, and then it can determine it is the optimal number of cloud market segments.

The mapping process is mainly determined by job priority, an average number of cores, and memory size, which is shown in Table 5—1. AMD [276] , Young [233], Michalski and Demiliani [314] and Feitelson [315] have provided some basic principles or guidelines to identify some common cloud application workload patterns. We assume the higher job priority, the critical workload is, such as SLA driven applications. If a workload has a lower job priority and consumes

sizeable computational power and memory, we assume it is a backend type of workload, such as Big Data Analytics applications.

Notice that Google's public dataset [178] only released the limited number of parameters for its cloud dataset so that the mapping of cloud applications is a rough estimation. If CSPs have their own operational datasets, the result should become more confident. As a result of market segmentation and business application identification, we can group the modeling process into three categories by three different analytic approaches. The roadmap of the modeling process is:

1. Define the utility functions of High Availability (HA) for segment 4 and Disaster Recovery (DR) for segment 5. This category is dependent on the specified SLA metric.
2. Build the utility functions for the data processing of Online Checkout (OC) and web hosting (WH) for segment 1, Virtual Desktop Infrastructure (VDI) for segment 3. This category is dependent on a response time
3. Model the utility functions of dynamic data processing (DDP) – dynamic content delivery (DCD) for segment 2 and backend (BE) workloads for segment 6. This category is dependent on the decision of risk.

5.2.3 **Utility Function for High Availability and Disaster Recovery**

HA business applications require mission-critical infrastructure or cloud resources. If we assume the downtime should be less than 5 minutes / per annum, then the service level agreement (SLA) must be higher than five-9s (or 99.999%). If this SLA is required, it means any failure of cloud infrastructure would lead to a catastrophic impact on the business revenue [235] for running business applications. One of the examples would be a Customer Relation Management (CRM) system, (e.g., Seibel), financial system, (e.g., stock trading), online banking platform, fast delivery ordering system, etc.

Suppose a firm hosts one of the HA business applications on cloud infrastructure is offered by a CSP. The cloud architecture supports the mission-critical applications shown in Figure 5—4. It means if one of the VMs fails, the workloads that are running on the faulty VM can be automatically transferred to another VM or a VM cluster (See Figure 5—5)

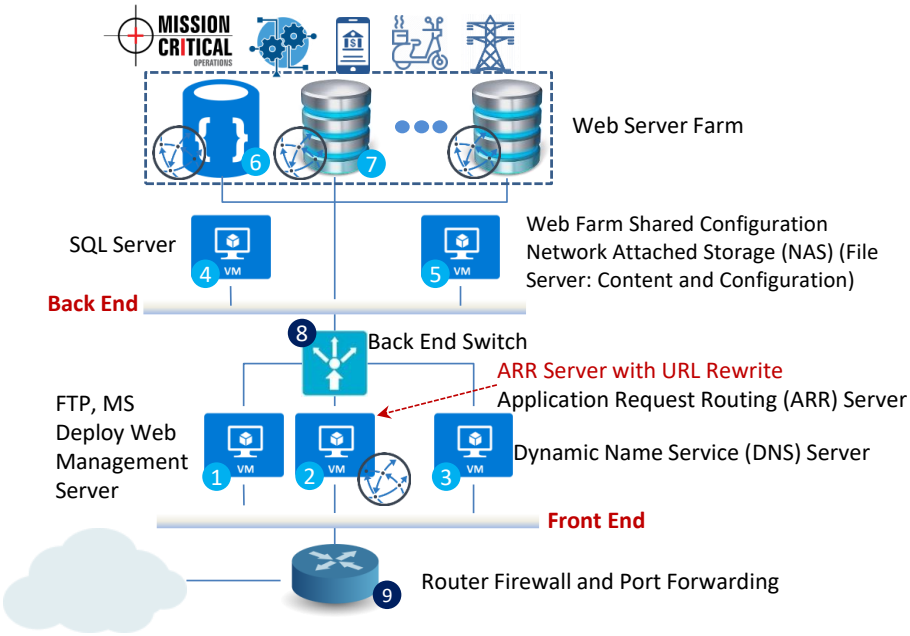


Figure 5—4 A Typical Web Hosting Architecture

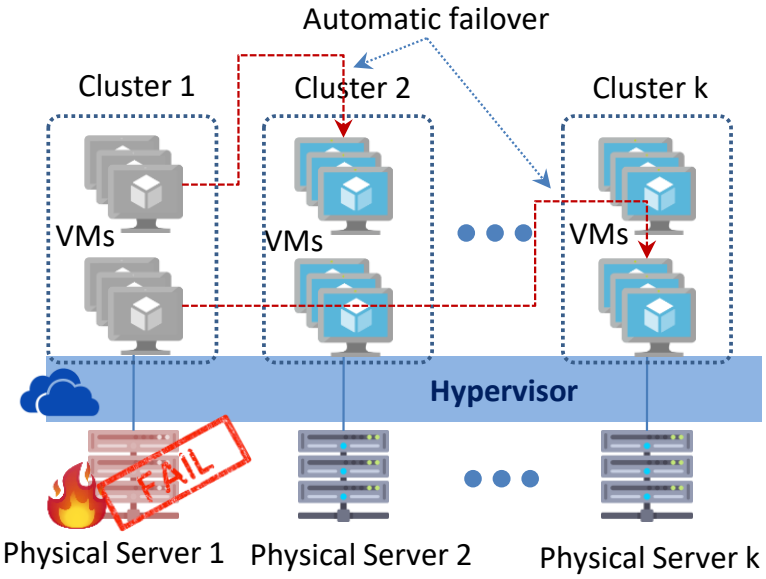


Figure 5—5 High Availability Cloud Infrastructure

If we assume the VM failure rate is μ (assume each VM is allocated in the different physical machines), its restoration rate is λ; the question is how many VMs (or a server farm) for CSP to support a mission-critical business application? Moreover, if we assume this server farm will impact on the business revenue for $π / per annum, how can we define the utility function for the cloud customer?

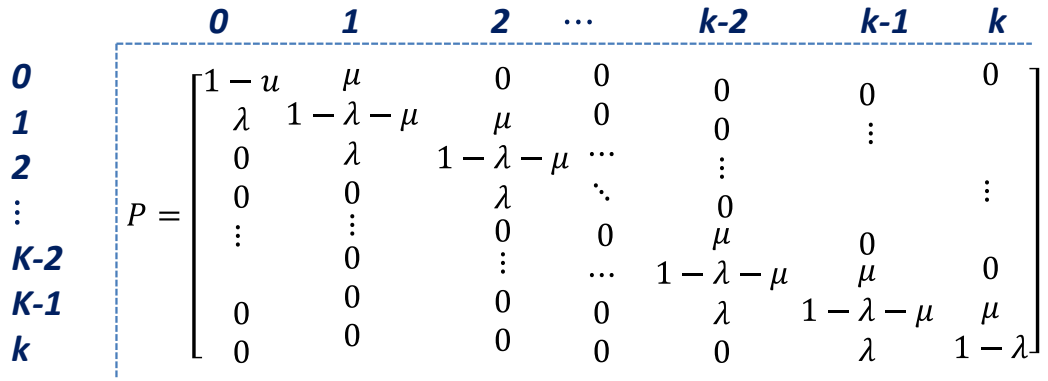


Figure 5—6 Markov Chain Diagram for Required "*k*" of Physical Servers

We can apply the Markov chain analysis to this problem. Assume we need "k" VMs to support the requirement of five 9s SLA. We can draw a diagram, as shown in Figure 5—6 based on the above assumptions. The k number of VMs can form a Markov chain system. This system is ergodic because we can verify the number of steps of the system would be exact "*k+1*" transitional states from any state to any other state. This means that the process can be characterized as a steady-state vector for a long run [219].

According to this Markov chain diagram shown in Figure 5—6, we should have a $(k+1) \times (k+1)$ Markov chain probability transitional matrix described in Figure 5—7. Based on this matrix, we can achieve a steady-state vector from Equation 5-1.

$$V_s = [V_1, V_2, V_3, \cdots, V_{k-1}, V_k] \tag{5-1}$$

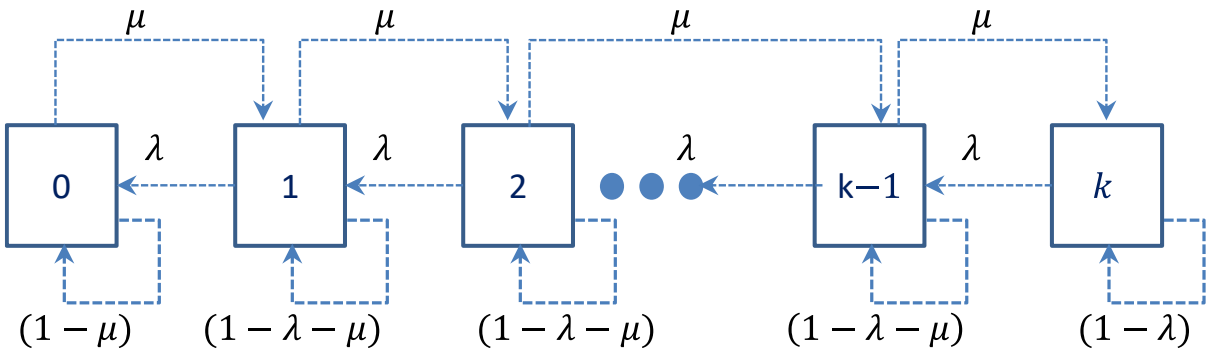


Figure 5—7 $m \times m$ Markov Chain Matrix

If we assume a failure probability μ of a VM or physical server [221] is 0.004 to and a faulty restoration rate λ is 0.2. Based on these assumptions, we can calculate the result of the steady-state vector via a transition probability matrix (number of VM failed) shown as the following equation 5-2

$$V = [0.98, \quad 0.0196, \quad 0.000392, \quad 7.84E-06, \quad] \tag{5-2}$$

The calculation result shows that the number of hot-standby VMs is at least three if we want the specified SLA is higher than seven 9s.

To generalize this transitional matrix, we can define a function V_k as a probability of kVMs is down. We want to find the minimum number of k such that the probability of this downtime is less than a specified time ϵ (e.g., five minutes /per annum)

$$V_{(k-1)} * \mu \leq \epsilon \tag{5-3}$$

Based on Figure 5—6 and Figure 5—7, if the system has a steady-state, we can derive an equation from 5-4 to 5-9 to calculate the k value.

$$V_i = V_0 \left(\frac{\mu}{\lambda}\right)^i, \qquad \alpha = \frac{\mu}{\lambda} < 1 \tag{5-4}$$

$$V_k = V_0 \alpha^{k-1} \mu \leq \epsilon \tag{5-5}$$

$$V_0 = \frac{1-\alpha}{1-\alpha^{k+1}}, \qquad \frac{1-\alpha}{1-\alpha^{k+1}} \alpha^{k-1} \mu \leq \epsilon \tag{5-6}$$

$$\alpha^{k-1} \leq \frac{\epsilon}{(1-\alpha)\mu + \epsilon\alpha^2} \tag{5-7}$$

$$(k-1)ln\alpha \leq ln\left(\frac{\epsilon}{(1-\alpha)\mu+\epsilon\alpha^2}\right) \tag{5-8}$$

$$k \geq \left\lceil 1+\frac{ln\epsilon - ln[(1-\alpha)\mu+\epsilon\alpha^2]}{ln\alpha}\right\rceil \tag{5-9}$$

Where any V_i is a probability distribution vector in the ergodic system, V_0 is the initial state of the probability distribution vector. V_i also indicates the probability that the system had i failures and now using the resource (i+1). $\epsilon = 1 - 0.99999$ (Five-9s: a specified SLA).

Note that equation 5-9 defines the probability transition from the $k-1$ state (the last VM) to the k state (or all VMs failure). The k value of equation 5-10 should be round up to the up ceiling that is not less than k. Once we have the result of k, we can define two customers' utility functions for market segment 4 and segment 5.

$$U_4(q) = \begin{cases} K_4, & 1 \leq q \leq k \\ 0, & k < q \leq q_m \end{cases} \tag{5-10}$$

where "K_4" is a revenue coefficient value, "q" is the variable of number of VMs, q_{m} is the maximum number of VMs (Refer to section 1.1). The interpretation of Equation 5-10 is that the cloud customers will purchase the maximum k number of VMs in segment 4 to meet their SLA requirements. This means that the "k" number of VMs will contribute the customers' business revenue together and each VM has the same utility value. For example, if one VM contributes business value is $1.0/per hour, then, the business customer has to purchase at least three VMs to run the business application, such as SQL database servers so that it can guarantee five 9s SLA delivery.

If the number of VM is higher than the required number k, its value of the revenue contribution will be diminished to zero. Therefore, the utility value is equal to zero shown in Equation 5-10. If a CSP's market strategy is to target Small Medium Enterprise (SME), then we can define the scaling coefficient K_i values by Equation 5-11

$$K_i = B_i / \left(\sum_{q\in[1,n]:U_i[n]\geq p^*} U_i[q] \right), \quad i = 1 \cdots S \tag{5-11}$$

where B_i is the annual revenue in each market segment of different categories of SME. p^* is the optimal price that is offered by CSP. "S" the maximum number of market segments, and "i" is a variable of the market segment.

For segment 5, it can also be considered as another type of mission-critical workload for business continuity because we have concluded this segment is to run DR or DR as a Service (DRaaS) applications. According to Luetkehoelter [316], the definition of DR is "the process of mitigating the likelihood of a disaster and the process of returning the system to a normal state in the event of a disaster." It can be a part of the HA workload, which is similar to a database backup. The difference is that DR mirrors everything in different physical locations. Therefore, the DR solution often requires a more quantity of VMs than the database backup application, but this requirement is dependent on a "likelihood" of DR.

Mathematically, this likelihood can be translated into a percentage of riskiness in terms of business impacts. This risk assessment [241] should be determined by a business continuity plan. We can formulate Equation 5-12 for the cloud customer's utility value.

$$U_5(q) = \theta K_5 \quad 1 \leq q \leq q_m \tag{5-12}$$

where θ is a potential risk rate (a percentage) to impact the cloud customers' revenue because a disaster occurs. It means that the cloud customers will only purchase the number of VMs when the price of VM (p^*) is below the specified threshold level of their utility value. If the VM price is higher than their utility value of a likelihood disaster, they will stop to purchase any cloud resource from CSPs and build the on-premises infrastructure. In addition to the mission-critical applications, the utility function for e-commerce can also be modeled by a Markov chain process.

5.2.4 **Utility for Queueing and Static Data Process**

E-Commerce applications, such as shopping cart, electronic data interchange (EDI), online catalogs, consist of a business processing module [244], which can be characterized as a type of queueing workload pattern [245]. One of the typical examples is the online checkout (payment) processing system shown in Figure 5—8 for an example of a web hosting service. It merely means that the end-users are lining up a queue for checking out due to online purchasing or ordering.

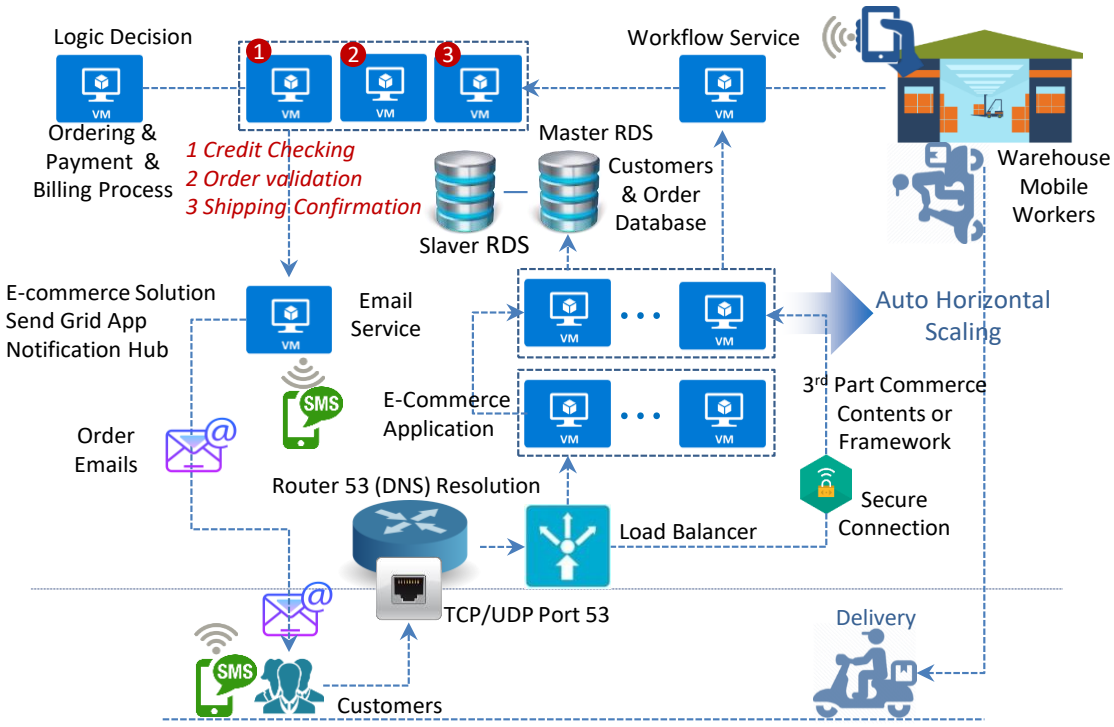


Figure 5—8 Typical Architecture of Checkout Application

Assume there is only one virtual machine (VM) that has been allocated to handle the end-users' checkout requests (λ_1) with a specified process capacity (μ_1), we would like to know how long (w_1) the end-users have to wait to complete the checkout process? We can use the simple M/M/1[17] to model [221] this process, which we can calculate out the expected waiting time for the end-users (online purchasers) from Equation 5-13.

$$w_1 = E[T] = \frac{\lambda_1}{\mu_1(\mu_1 - \lambda_1)} + \frac{1}{\mu_1} = \frac{1}{\mu_1 - \lambda_1} \tag{5-13}$$

where T is the total expected time for an end-user within the checkout system, which includes the waiting time to be processed for checkout, this expected waiting time is critical for the cloud business customer who runs an e-commerce business (or an online shop). If the time is too long, the end-user will start to lose patience and just merely switch to another portal (online) shop at the click of a finger. In other words, the expected waiting time will impact the cloud customer's

[17] Kendall's Notation of a Queueing System, A/B/C, A indicates the inter-arrival time distribution, B indicates the service time distribution, and C indicates the number of servers.

e-commerce business revenue. On the other hand, if we allocate too many VMs resources to the checkout system, many VMs will stay idle. It will increase cloud customers' operation expenditure (Opex). The issue for a CSP is how to establish an adequate utility function to model the hosting business value for its cloud customers.

Taking an example, if we assume the average arrived rate of the end-user as $\lambda_1 = 8$ /per hour, and $\mu_1 = 10$/per hour [241], the expected average waiting time will be 24 minutes in the queue. If we include the average 6 minutes of the processing time for checkout (payment), a random end user will spend a total of an average of 30 minutes in the system shown in Equation 5-14.

$$w_1 = \frac{1}{\mu_1 - \lambda_1} = \frac{1}{10 - 8} \times 60 = 30\ minutes \tag{5-14}$$

Based on our own experiences, 24 minutes of queueing time would be unacceptable for an online business. To reduce this expected waiting time, we have two possible solutions: one is the vertical scaling, which is to increase the VM's capacity μ_1 by selecting large capacity VM so that the time of the checkout process can be reduced. For example, if we double the VM capacityμ_1= 20/per hour, the waiting time w_1 can be reduced to 5 minutes. The other solution is the horizontal scaling that is to add more VMs with the same capacity into the checkout system, which can also decrease the queueing time w_q. If this is a case, the problem of M/M/1 becomes an M/M/s [221] model, which can be described in Figure 5—9.

If the workload of e-commerce application is highly fluctuant, then the horizontal scaling would be a preferred solution. It also adds a bonus of the high availability into the system, which we illustrated this point in the previous section. Moreover, the different end-user might have different lengths of responding time to the checkout system. For example, a new end-user may take more time to respond to the checkout system than a frequent user.

If we select a horizontal scaling solution, then Erlang's delay formula [221] can calculate both queueing and the total processing time ($w_q\ and\ w_s$) for the number of VMs required.

$$w_q = \frac{\alpha^s p_0}{s!\, s\mu_1 (1 - \rho)^2} \tag{5-15}$$

$$p_0 = \left[\sum_{k=0}^{s-1} \frac{\alpha^k}{k!} + \frac{(\alpha)^s}{s!}\left(1 - \frac{\alpha}{s}\right)^{-1}\right]^{-1} \tag{5-16}$$

$$\alpha = \frac{\lambda_s}{\mu_s} < 1, \qquad \rho = \frac{\alpha}{s} = \frac{\lambda_s}{s\mu_s}, \qquad w_s = w_q + \frac{1}{\mu_s} \tag{5-17}$$

where $w_q$is the queueing time for the end-users in the queue to be served. w_s is the processing time in the checkout system. "s" is the number of VMs required to reduce the queueing time for the end-users. "k" is a variable of VM.

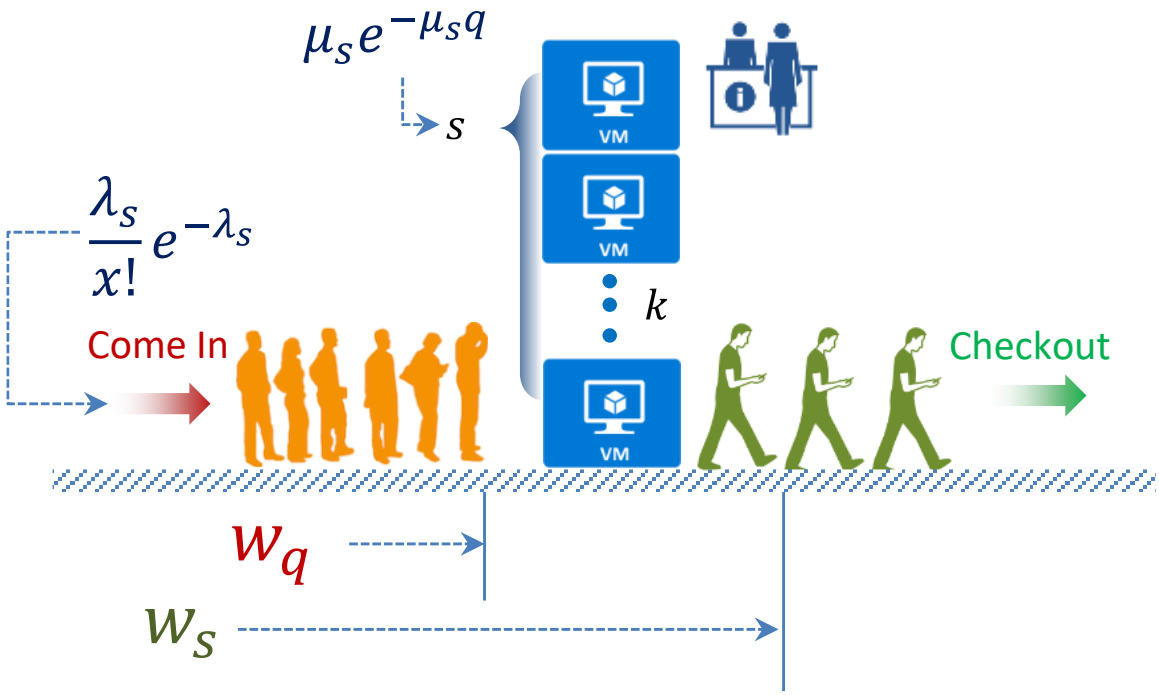


Figure 5—9 M/M/s Queueing Model

Using the same $\mu_1$and λ_1 as the M/M/1 model, we should have the following calculation results in Table 5—2.

Table 5—2 Calculation Results for M/M/S model

No of VMs	ρ	p_0	w_q (minutes)	$1/\mu_1$ (second)	w_s (second)
1	0.800	1	24	360	1,800
2	0.400	0.4285714	1.142857	360	428.6
3	0.267	0.4471545	0.141907	360	368.5
4	0.200	0.5020080	0.02008	360	361.2
5	0.160	0.5392432	0.002504	360	360.2

If we plot out the result of queueing time against the incremental number of VMs shown in Figure 5—10, we can have an approximate trend line in a power function. According to both Table 5—3 and Figure 5—10, we see that queueing time decreases sharply after the 2nd VM or 3rd VM. Therefore, we can use quotation 5-18 to approximate a utility function for market segment 1:

$$U_1(q) = K_1 q^{-c}, \qquad 1 < q < q_m \tag{5-18}$$

where K_1 is a scaling coefficient. "c" is a constant that is to determine the gradient of the power equation. q_m is the maximum quantity of VM that the customers of segment-1 may purchase.

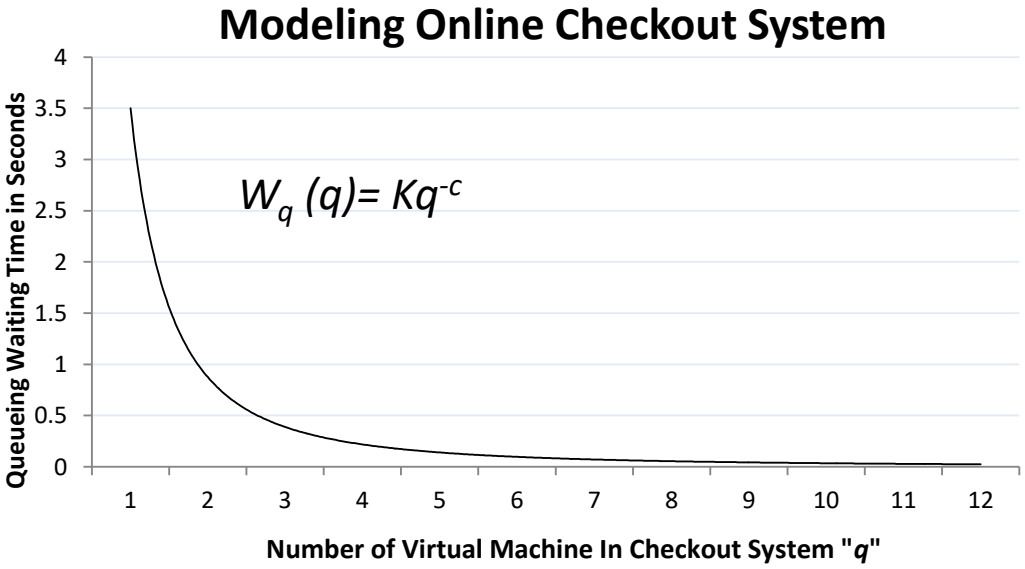


Figure 5—10 M/M/s Queueing Model

On the other hand, if the CSM needs reducing the overall processing time (both queueing and processing time), the cloud customer or the e-commerce business owner has to have a solution of combining both vertical and horizontal scaling.

If the λ_s value is relatively small in comparison with μ_s, the power function is sufficient to model the customer utility value. If the λ_s the value becomes larger, then adopting the discrete function (Equation 5-10) is a good idea to descript the cloud customer's utility value because a guaranty to deliver SLA becomes a major issue when the average number of end-users increases.

Alternatively, we can also use a linear function [236] to provide a solution when there is a constant rate of changing in terms of VM demand and utility value. This topic leads to our next

issue of how to model customer utility in the segment-3. The workload characteristics of this segment have been classified as "Virtual Desktop Infrastructure (VDI)." There are many VDI performance metrics of a hosting environment regarding users' experiences, such as the peak of Input/Output Per Second (IOPS), storage capacity, response time, Read/Write ratio, future growth, etc. If we assume these metrics have been prefixed during the Proof of Concept (POC) period before VDI rollout, the additional VM will only add Opex and a burden to the cloud customers. So, we can use a linear model to measure the cloud customer utility value because an end user's response time is calculated as a linear model based on the cloud resource request [245] [246].

$$U_3(q) = K_3(rq + q_m), \qquad r < 0 \tag{5-19}$$

where "r" is a constant, but it is negative to reflect the economic principle of the diminishing return.

5.2.5 **Utility Function for Backend and Dynamic Data Processing**

When we encounter backend and dynamic data processing types of workload, such as dynamic content (optimized dynamic content) delivery, clone server, Network File Sharing (NFS), and cache proxy, we should use different mathematical models to measure the cloud customer's utility values in term of the end-users' experiences. According to [80], we can use Equation 5-20 to model the customers' utility value for the dynamic content workload for market segment 2. It measures the constant relative risk aversion (CRRA) when the cloud customer is facing some uncertainties.

$$U(q) = \begin{cases} \dfrac{q^{1-\alpha} - 1}{1-\alpha}, & \alpha \in (0,1) \\ ln(q), & \alpha = 1 \end{cases} \tag{5-20}$$

where "α" is to measure the degree of relative risk aversion. Based on the Pratt-Arrow absolute risk aversion function (Equation 5-21 R_r), we can measure the absolute value of risk aversion, which is to define the coefficient value at "q." R_r is a negative exponential (or inverse) function at "q" when α is greater than one

$$R_r = -\frac{U_2^{"}(q)}{U_2'(q)} = \frac{dU_2'(q)}{dq}\frac{q}{U_2'(q)} = \frac{\%\Delta U_2'(q)}{\%\Delta q} = (1-\alpha)q^{-1} \tag{5-21}$$

Practically, it means that if R_r is decreasing with respect to VM quantity "q," the cloud customer will be less sensitive towards risk aversion when the number of VMs is increasing.

We can also use the exponential utility function to model the backend type of workload for market segment 6. The exponential function gives us the value of constant absolute risk aversion (CARA) (Refer to Equation 5-22 and 5-23)

$$U_6(q) = K_2 \begin{cases} \dfrac{(1 - e^{-\alpha q})}{\alpha}, & \alpha \neq 0 \\ q, & \alpha = 0 \end{cases} \tag{5-22}$$

$$R_a = -\frac{U_6^{"}(q)}{U_6'(q)} = \alpha \tag{5-23}$$

where α represents the constant absolute risk aversion [222]. When $\alpha = 0$, it means risk neutral, and when α<0, it is risk-seeking. In this chapter, we set the value of $\alpha < 0$.

The size of the backend workload is often quite large, and the processing environment is complicated because it involves different issues of cloud architecture, planning, and resources scaling, e.g., database replication (1:1 replication of both master and slave for zero-downtime), read replica (use the ave as a read-only instance), in-memory caches (Key-Value Store for the session and state data, across cloned instances), and etc. As a result, we can set the α value either less than zero or equal to zero to estimate the customers' utility values. In other words, we use the exponential function with α<0 to describe a customer's utility values in terms of acquiring VM resource because the customers are sensitive towards a cloud price and the backend workload can be interruptible.

5.2.6 **Define the Coefficient Values**

The final issue is how to determine the value of K_i and α. The scaling coefficient of K_i is dependent on the business revenue or profit that a particular type of VM instance (such as AWS's extra-large instance) can help cloud customers to produce. For example, if we target the average profit of SME is around $48K-$110K/per annum [223], we can approximately estimate the profit for each VM to generate is between $0.95 and $1.9/per hour [249] for various cloud applications across six market segments.

It is challenging to determine the value of risk aversion α because it measures the cloud customers' subjective feelings when they are facing uncertain outcomes [247]. Based on [246] and [249] recommendations, we set the value of risk aversion is equal to 0.4 in this research.

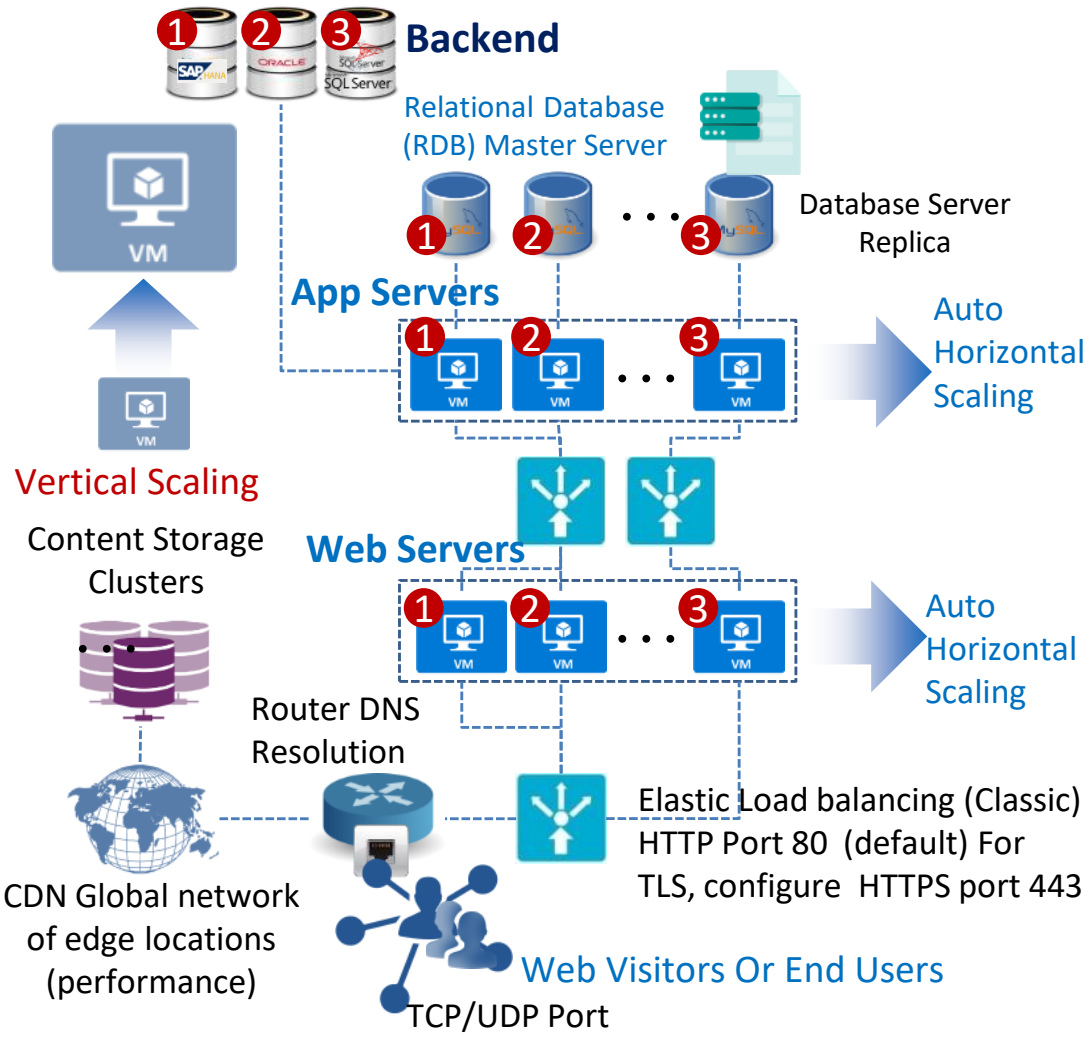


Figure 5—11 Typical Architecture of Web Application Hosting

According to all the above coefficient values, we can normalize the average utility value up to $1.50 (per/hour), and the minimum utility value is equal to $0.00 across all six market segments. The maximum number of VM is an arbitrary number. Here, we set to q_m=12. It is just a matter of a scale. Figure.5-11 illustrates the number of web applications may run on a single cloud platform. As a result, the quantity of q_m could be various from one case to another. In other words,

if the solution architecture is changed, the value of $q_m$will also be changed. For example, if the business requires running different types of database, the solution architecture should be altered.

Figure 5—11 also shows an example of an architecture solution for cloud resource scaling, which can be either horizontal or vertical. The decision of cloud resources, whether it should be vertical or horizontal scaling, depends on the definition of customers' business requirements, such as CSM.

Finally, the detail values of multiple utility functions can be constructed in Table 5—3, which is based on the segmented market. It provides a solution to the problem that is raised in section 1. Table 5—6 is a foundation to generate different cloud price models. CSP can leverage various price models to identify an optimal price point of each model for its profit maximization.

Table 5—3 Cloud Customers' Utility Table

VM No.	Seg 1	Seg 2	Seg 3	Seg 4	Seg 5	Seg 6
1	$1.50	$0.00	$1.50	$1.50	$0.75	$0.01
2	$0.75	$0.23	$1.36	$1.50	$0.75	$0.02
3	$0.50	$0.41	$1.23	$1.50	$0.75	$0.03
4	$0.38	$0.57	$1.09	$1.50	$0.75	$0.05
5	$0.30	$0.71	$0.95	$1.50	$0.75	$0.08
6	$0.25	$0.84	$0.82	$1.50	$0.75	$0.13
7	$0.21	$0.97	$0.68	$0.00	$0.75	$0.19
8	$0.19	$1.08	$0.55	$0.00	$0.75	$0.29
9	$0.17	$1.20	$0.41	$0.00	$0.75	$0.44
10	$0.15	$1.30	$0.27	$0.00	$0.75	$0.67
11	$0.14	$1.40	$0.14	$0.00	$0.75	$1.00
12	$0.13	$1.50	$0.00	$0.00	$0.75	$1.50

5.2.7 **Summary of Modeling Multiple Utility Method**

Up to this point, we have generated all six utility functions along with six cloud market segments, which can be summarized in Table 5—4. This table covers multiple utility functions with different cloud customers' preferences for various business applications. The pre-condition of the modeling utility function is the result of cloud market segments. The number of market

segments is derived from a CSP's cloud business strategy and targeted cloud customers. These market segments can be analyzed by three types of analytic approaches.

A CSP can also have more or less than 6 market segments. According to [238], the suggested number of the market segments is between 5 and10. Overall, the number is dependent on a portfolio analysis to meet the CSP's business objectives by balancing sales growth, capital investment budget, cash flow, cloud technology expertise, and business risk. For example, if the CSP would like to explore a particular niche cloud market (e.g., cage-level physical security), a customer's utility function may be defined differently.

Table 5—4 Cloud Customers all Utility Functions

Business Application workload	Analytic Approach	Market Segment	Cloud Customers Utility Function
Online Checkout	Queuing Theory	1	$U_1(q) = K_1 q^{-c}$
VDI		3	$U_3(q) = K_3(q_m + rq), \quad r < 0$
High Availability Data	Markov Chain Analysis	4	$U_4(q) = \begin{cases} K_4, & 1 \le q \le k \\ 0, & k < q \le q_m \end{cases}$
Disaster Recovery		5	$U_5(q) = \begin{cases} \theta K_5 & 1 \le q \le \mathrm{k} \\ 0 & k \le q \le q_{\mathrm{m}} \end{cases}$
Dynamic Content Delivery	Risk Assessment	2	$U_2(q) = K_2 \begin{cases} \frac{q^{1-\alpha} - 1}{1-\alpha}, & \alpha \in (0,1) \\ ln(q), & \alpha = 1 \end{cases}$
Backend Data Processing		6	$U_6(q) = K_6 \begin{cases} \frac{(1 - e^{-\alpha q})}{\alpha}, & \alpha \neq 0, \alpha < 0 \\ q, & \alpha = 0 \end{cases}$

When we estimate K_i coefficients, we balance the values of all the scaling coefficients to be equivalent by similar grouping revenue of SME together. If a gap of the coefficient value is too large, then the higher value of the coefficients would have more influence on the optimal price of a VM. To visualize all utility functions of Table 5—4, we can plot out all six cloud customer utility functions along with the number of VM variations in Figure 5—12. As we can see, the blue line represents the utility value, and the red line captures the marginal utility.

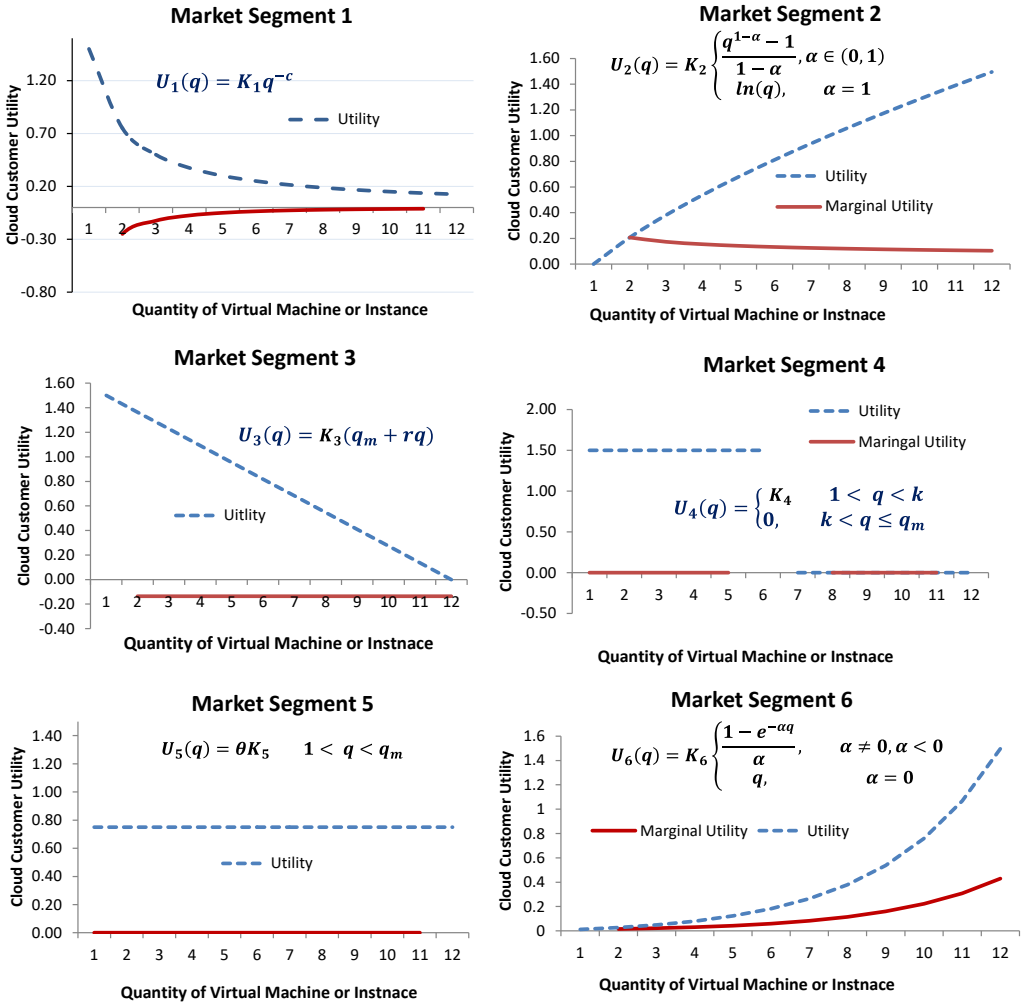


Figure 5—12 Six Cloud Utility Functions for Six Cloud Market Segments

We argue our method of modeling multiple utility functions is dependent on both internal (strategic objectives, cost, expertise, cash flow, targeted customers, etc.) and external (CSM, cloud customers' revenue or profits, and market segments) rationalities for a CSP to achieve the maximum profit by identifying the optimal price. Our basic idea of modeling the cloud customers' utility functions is to assign SME customer's revenue into each VM that can generate for the cloud customer, which is the concept of value co-creation [239] [10]. In order to compare with different modeling methods, the following section is going to survey previous different methods.

5.3 **Related Work**

The modeling customer utility functions for various hosting services can be traced back to the beginning of the dotcom-booming era. Doyle et al. [206] [232]proposed a model-based approach to optimize the hosting of hardware resources for the specified SLA. The goal of their work is to demonstrate how to provision the server resources for web hosting applications effectively. Although the paper adopted the term "utility" and made good progress in hosting service modeling, the real meaning of utility is the usefulness of server functionality rather than an economic sense of subjective measurement for customer satisfaction. Similarly, Appleby et al. [207] proposed an SLA-based management system, which is named "Oceano" for e-business. It is based on a set of predefined metrics that consists of seven parameters. Their approaches can be considered as a policy-based scheme for computer resource allocation. The policy mainly reflects what a provider wants rather than the explicit measurement of the customer's satisfaction and experiences.

In contrast to Appleby, Walsh et al. [208] gave an explicit measurement for the customer's utility functions in order to automate the computer resource distribution. The utility functions are an autonomic scheme to manage web hosting workloads running on a Linux cluster. They defined the utility $U(S, D)$ as a function of two independent variables: service level (S) and current demand (D), which is measured by an average of forecast demand (D'). S is a function of the other three independent variables that are control parameters (C), which is responsible for optimizing the utility $U(S, D)$, current resource level (R), and demand (D). Overall, the customer utility value can be estimated by variables of C, R, and D' and defined as Equation 5-1 if the service performance (S) is specified.

$$\widehat{U}(R) = \max_{C} U[S(C, R, D'), D'] \tag{5-24}$$

This service performance is designed to run the application of IBM WebSphere and DB2. The paper argued that the utility $\widehat{U}(R)$ is defined by a sigmoid function in terms of the average response time. They did some pioneer works regarding utility functions. However, the authors left the details of modeling the sigmoid function. In comparison, Bennani et al. [209] gave some details for their proposed utility function in the sigmoid form (Equation 5-25) regarding online application environments.

$$U_{i,s}\left(R_{i,s}, \beta_{i,s}\right) = \frac{K_{i,s} e^{-R_{i,s}+\beta_{i,s}}}{1 + e^{-R_{i,sthe}+\beta_{i,s}}} = \frac{100 e^{-R_{i,s}}(1 + e^{\beta_{i,s}})}{1 + e^{-R_{i,s}+\beta_{i,s}}} \tag{5-25}$$

where, $K_{i,s}$ is a scaling coefficient. $U_{i,s}$ is the function of $\beta_{i,s}$, $R_{i,s}$, $R_{i,s}$ is the response time for "i" type of application environment (equivalent to a type of workload) with "s" type of classes of transactions (equivalent to a type of virtual machines), $\beta_{i,s}$ is desired or targeted SLA (equivalent to the customer performance metrics). The goal of their paper is to come up with a solution (or global with controller) that can automatically assign different types of workloads to the adequate size of the server for the data center infrastructure. The value of their utility function varies between 0 and 1. Its scaling coefficient is corresponding to the upper bound of the throughput of the job or workload completion for a certain application. The authors assume the higher throughput, the higher utility value is. From this perspective, the meaning of utility has become "utilization" or "utilization" rate of IT resources.

Following a similar line of reasoning, Kephart et al. [210] proposed a self-management system that is based on the utility framework in order to achieve resource efficiency in a prototype of a data center. The value of the utility function is between -1 and 1. The independent variables of the utility function can be either response time or the number of physical servers. Menache et al. [211] further developed this idea and proposed a long-term solution for cloud computing resources in terms of maximizing the social surplus, which is then aggregated individual user's utility of executed jobs minus workload-dependent operation expenses (Opex). This social surplus is equivalent to Bennani's [209], the global controller. The individual utility function (or local controller) of each user is presented in Equation 5-26.

$$U_i(z_i) = V_i(z_i) - Pz_iT_i(z_i) \tag{5-26}$$

where, $V_i(z_i)$ is the value that user (i) assigns to executing job required z_i amount of resource for P unit price for mean service time, $T_i(z_i)$. Although it was just a theoretical discussion, their work was the first time to define the utility in a microeconomics sense. The paper made a good contribution to the utility function definition. However, some assumptions need to be further consolidated. For example, the assumption of M/G/∞ model means no resource restriction. If this is a case, the optimal solution will become impracticable. Just as the authors highlighted, their analytic model only provided a convenient starting point for future research topics of cloud computing, such as revenue, profit, and pricing. Nevertheless, the paper indicated there is an optimal point by a linear usage based-tariff (or fixed price/per unit resource/ unit time).

Weintraub et al. [212] presented a survey plus ranking (ordinal) model that is a conjoint analysis (ranking multiplied by the weighted coefficients) that shows how to maximize the user's total

utility from a set of cloud services offered by CSPs. This utility model is the cloud feature or characteristics-based services for selecting a preferred CSP. It is a utility model in terms of customer preference choice.

Regarding the preference choices, Burda et al. [213] examined consumer preferences for the cloud archiving services from a student's perspective. Burda et al. adopt conjoint analysis (a survey-based statistical technique) to quantify customers' utility levels in three market segments based on the customers' demographic parameters, such as age and gender. Although the authors' did not explicitly adopt the term of market segmentation, the paper was the first time to introduce the idea of market segmentation. Their study focused on business to consumer (B2C) market rather than business to business (B2B) market.

Minarolli et al. [214]adopted a similar approach as Bennani et al. [209], which is to set up both local (like a transponder) and global controllers (or a central management system) to allocated cloud resources pool. They defined the utility function is a simple linear Equation 5-27, which is also the extension works of Walsh [208] and [215]

$$U_i = \alpha_i \cdot S_i \tag{5-27}$$

where the amount of dollar α_i is paid per unit of CPU resource, and S_i is the located CPU resource to VM_i or shared physical CPU utilization. However, this resource consumption model is just one of the utility functions if the cloud customers take a natural risk attitude. One of the critical issues of their work is the meaning of utility was not clearly defined in the economic sense from a customer's perspective. The work described the term utility as optimizing the cloud resource pool. The unit of the utility measurement is switched between the quantity of VMs and the length of response time.

Similarly, Garg et al. [216] [217] provided an admission control solution for a similar problem, but they described the term of utility as a resource scheduling rather than an economic sense of the utility function. The assumption of their application is the non-interactive or static workload. Their solution is to achieve the optimizing scheduling between the specified Quality of Service (QoS) requirements and resource provisioning.

In comparison with others, Chen et al. [69] made some contributions to the utility function defined in terms of microeconomic sense. They presented scheduling solutions from a cloud customer's utility perspective. They showed that cloud customers could effectively bid for

different types of VM spot instances under the conditional metrics of profits, customer satisfaction, and cloud resource utilization. The detail of the cloud customer utility function is shown as follows:

$$U(p,t) = U_0 - \alpha p - \beta t \tag{5-28}$$

where, $U_0$is the maximum utility that the service delivery to the customers. It is proportional to the size of the service request. Both α and β is the coefficient of price "p" and response time "t." Again, the cloud customer utility value is a linear function of price and response time. The application of their utility function is based on running the x264 application for video scripts' encoding and decoding tasks. Overall, we can highlight the main contributions and gaps of each modeling method for the customers' utility function over the last two decades in Table 5—5.

Table 5—5 Summary of All Methods of modeling Utility Function

Modeling Methods	Solution Idea	Pros	Cons	Application
Model-based	Slices of computer resources, and time for resource management	Enable a provision of multiple resources in an interactive way	It only works for the prototype. The concept of utility is not the economic one	Web-based service or Content Distribution Network
SLA-based	Leveraging SLA to allocate resources	Dynamic, flexible, scalable resource allocation	The resources assumption has no limit. It is not the real economic utility	e-business hosting
Resource-based	Utility function to allocate resource	Self-optimization of computing capability	A data center management system rather than a model for the utility function	Data Center, CND, Video streams
Social Surplus-based	Adopting Social Welfare idea and leveraging queueing theory	The social effects of using cloud resources	Just a theoretical model. Similar to global and local controls	Academic discussion to prove the convexity assumptions
Empirically Calibrated	Empirically calibrated model	Provider an alternative way of utility modeling	Limited applications remain empirical	Intend to explain major cloud leaders' market behaviors
Price-Quality	Single and multi-tiered solution	Define the price-quality from NE perspective	Only apply it for a special case under particular assumptions	Theoretical interpretation
Capacity-Aware	Non-additive utility function	Dynamic	Mixed with users and CSP utilities. Pre-negotiation of deliverable SLA	Negotiable cloud resources or Grid computing
Conjoint Analysis	Three layers of customer utility	Survey plus ranking	Too arbitrary to build the utility function, not very explicit.	Cloud Customers Survey data
Framework-based	Utility function policies	Two types of Integrated Utility values	Prototype, impracticable	Data Center Environment
Simple Linear	A two-tier resource management approach	The balance between QoS and Operation Cost	Confusing with CSP and Customers Utility	VM resource Allocation

As Table 5—5 indicates, the majority of previous works are more like a resource management scheme with the aim of managing cloud resources. Strictly speaking, many models did not consider cloud market segmentation. The term "utility" was not defined as a subjective measurement unit from a cloud customer perspective. The meaning of utility was often swinging between the supply and demand sides.

In contrast, our utility modeling process is driven from three aspects: define cloud business customers' service metrics, model the utility functions based on the assumption of the cloud market segmentation theory and focus on customers' revenue and profit contribution. The following section gives a full comparison of our modeling method and other methods.

5.4 Performance Evaluation

The performance evaluation is divided into two parts. The first part is to compare the market share value. The second part is to compare all economic values, which include revenue, profit, an optimal price, and a marginal cost based on the same price model of "on-demand."

5.4.1 Comparison of Cloud Market Share

In comparison with some previous modeling methods of the utility functions (Refer to Table 5—6 for details comparison), our modeling approach has the following advantages: First, the measurement unit of all the utility functions is unified under the customer's revenue or profit (dollar). Second, this measurement is tangible and can be compared. Third, the different market segment has different types of utility functions. Fourth, each market segment is associated with one type of cloud business application. There are a total of six market segments. It avoids "one size fits all." Fifth, we only model the cloud customers' utility functions in this chapter. Sixth, In contrast to previous SLA studies, we clearly specified the number of VMs required generating cloud customer's revenue. Seventh, we lay out a clear definition of utility in upfront to avoid any misinterpretation.

Table 5—6 Performance Evaluation of Different Methods of Utility Modeling

Methods of Utility Modeling	**Utility Functions**	**Cloud Market Share**	**Measurement of Utility Unit**	**Ind. Variables of the Utility function**
Proposed Method	$U_i(q),\ i = 1 \cdots S,$ $q = 1 \cdots q_m$	**83~ 100%**	Customer's Revenue & profit	VMs
Model-based	$U(H) = \frac{1 - M^{\alpha}}{1 - T^{\alpha}}$	16%~17%	Hit Rate	Memory "M" & workload (object "T")
SLA metrics	$\hat{U}(R) = \max_{c} U[S(C, R, D'), D']$	16%~17%	Service Level	Control "C" Resource "R", Demand "D"
Resource-Based	$U_i = \alpha_i S_i,$	16%~17%	Performance metrics (e.g., response time)	A throughput of response time $R_{i,s}$, specified time $\beta_{i,s}$
Social Surplus-Based	$U_i(z_i) = V_i(z_i) - P z_i T_i(z_i)$	16%~17%	The expected resource within time and price limit	Price P, resource z_i Execution Time T_i Expected Value V_i
Empirically Calibrate	$U_{ijk} = w\left(v_i - \frac{c_i + 2^{k-1} p_{j1}}{2^{k-1} \alpha_j^{k-1} q_{j1}}\right)$	16%~17%	Expected value v_i minus combination of three variables	Price p_{j1}, delay time sensitivity c_i, quality level q_{j1}, workload w
Price-Quality	$U(pr, s) = P_i(pr, s) = \lambda_i(pr_1 - c_i - \rho_i) - \frac{\rho_i}{\overline{rt} - s_i}$	16%~17%	Payoff (resource request capacity) vs Price	CSP Price pr, response time s, unit Opex c_i and unit Capex ρ_i
Capacity Aware	$U(P_{N+1}, T_{N+1}) = \sum_{i=1}^{N+1} U_i^{Q} - \sum_{i=1}^{N} U_i^{R=\{SLA\}}$	16%~17%	CSP's profit $U(P_{N+1}, T_{N+1})$	Service Price P_{N+1} and response time T_{N+1}
Conjoint Analysis	$U(R_i, p_i) = \sum_{i=1}^{n} R_i p_i$	16%~17%	Preference ranking	Attribute ranking R_i & weight p_i
Framework Based	$U(f) = -e^{-5e^{-0.5 * f(A,E,R)}} + 1$	16%~17%	Sigmoid function value	Specified scenario parameters A, E, R
Simple Linear	$U(p, t) = U_0 - \alpha p - \beta t$	16%~17%	Service request satisfaction level	VM price "p" & response time "t"

The calculation of the market share is dependent on the assumptions of the number of market segments. If the model assumes the cloud market has a single market, the pricing model will address smaller proportional customers. For example, if AWS offers one price only for all its cloud customers, such as spot instance price, the majority of business customers will not purchase its cloud service because this cloud service cannot provide the service guarantee for some mission-critical applications. Therefore, a price model can only capture 16% ~ 17% (1/6) market shares in comparison with an assumption of six market segments. It is self-explanatory.

The value to capture more market shares only gives one indication for an addressable cloud market, but the ultimate purpose of market segmentation is to set up a pricing foundation for CSP to achieve profit maximization. Therefore, it is essential to validate our method through an experiment based on a particular price model –"on-demand."

5.4.2 Economic Values Comparison

The full details of price modeling can be found in the subsequent research work (See the Following chapter), which has been highlighted in Figure.5-1 for steps 3 and 4. In this work, we only give brief information on how we implemented our experiment through a particular price model and then we demonstrate the experimental results with different methods. Finally, we provide a comparison for different modeling methods (Refer to Table 5—7) for the justification of our claims.

5.4.2.1 Process of Evaluation

As we indicated, we adopt the "on-demand" price model for cloud pricing to implement our experiment. This price model is offered by nearly every leading CSP. This price model reflects one of the cloud characteristics, pay as you go (PAYG). The price model is value-based (Customer Surplus value-based), the "on-demand price model can be defined as

$$q_i[p] = q_i : \max CS_i[p] = \left(\sum_{j=1}^{q} U_i[j]\right) - pq \geq 0, \quad Q(p) = \sum_{i=1}^{S} q_i[p] D_i[p], \quad i = 1, \cdots, S \qquad (5\text{-}29)$$

where S is the number of market segments, which is equal to 6 from the above-market assumption. The q_i is a number of VMs to be acquired by the cloud customers in the market segment "i." This acquired quantity is determined by the maximum customer's surplus-value $CS_i[p]$ that is greater than zero for the given price p which is offered by a CSP, based on a defined utility function $U_i[j]$ which represents the external rationality (Refer to both Table 5—3 and Table 5—4.) for the "i" market segment while j is a variable of VM between i and q. q_i is a dependent variable of a price p. It means if the cloud price is changed, a quantity variation of each market segment will also follow (Table 5—7).

Table 5—7 Experiment Results of Comparison Uniform and Six Market Segment

Comparison for "on-demand" price model	Resource-Based With Single Market $U_i = \alpha_i S_i$, $i = time$	Simple Linear with Single Market $\boldsymbol{U(p,t) = U_0 - \alpha p - \beta t}$	Proposed Method with six market segment $\boldsymbol{U_i(q)}$, $\boldsymbol{i = 1 \cdots 6}$	Δ of Proposed Method and Resource-Based	Δ of Proposed Method and Simple Linear
Optimal Price	$0.751	0.955	$0.749	-0.27%	-21.57%
Unit Cost	$0.554	0.942	$0.274	-50.54%	-70.91%
Total Sales Vol.	4,760	1,859	5,920	24.37%	218.45%
Total Revenue	$3,435	$1,775	$4,440	29.26%	150.14%
Total Cost	$2,537	$1,751	$1,625	-35.95%	-7.20%
Total Profit	$898	$24	$2,815	213.47%	11,629.17%

If the cloud customer's surplus has been quantified, the maximum profit $\pi[p]$ can also be achieved by identifying the optimal price p^* (See Equation 5-30). Based on microeconomics, the profit equation can be easily defined as the total revenue $p * Q(p)$ subtracts the total cost $C[Q(p)]$ shown as following Equation 5-31.

$$p^* = arg\max_p \pi[p] \tag{5-30}$$

$$\pi[p] = pQ(p) - C[Q(p)], \quad c_u[Q(p)] \le p \le M, \quad c_u[Q(p)]\, Q(p) = C[Q(p)], \tag{5-31}$$

where $Q(p)$ is the summation of $q_i[p]$ of VMs multiplied by the estimated forecast for market demand (or predicted sales volume) $D_i[p]$ (See Table 1) of each market segment. M is the normalized maximum utility value. $c_u[Q(p)]$ is the unit cost, which represents CSP's internal rationality.

5.4.2.2 Dataset

The experiment dataset has already been presented in Table 5—3. To optimize Equation 30, we adopt the genetic algorithm to run our experiment. There are a number of software applications that can be applied to implement a genetic algorithm, such as Matlab, R and even Microsoft Excel Solver. The R package has two convenient packages: GA and Genalg, that can deliver quick results.

5.4.2.3 Experiment Results

If a CSP assumes the cloud market has only a single market with one definable utility function (e.g., either resource-based or simple linear utility function), the price of "on-demand" can only achieve either $898 or $24 profit respectively (shown in Table 5—7). In comparison with six

market segments with multiple utility functions, the profit margin can reach $2,815. In other words, our method of defining utility function can achieve 213% more profit than the resource-based method and 11,629% more profit than a simple linear while the unit cost drops 50% (resource-based) and 71% (simple linear) respectively.

Notice that the optimal prices of resource-based and our multiple utility functions are not much different, but the profit margin is 213% apart. This is because not all customers' utility functions are continuous. Some utility functions are discrete. If the evolution of different charging prices is plotted out for the on-demand price model, we can see there is a sharp drop in revenue and profit while the unit cost increases dramatically beyond the optimal price for the multiple-utility functions (more details in next chapter). This is similar to many retailers to adopt a psychological discounting price, such as $0.99 instead of $1 to boost sales volume or to increase their revenue in the retail industry.

We have validated our proposed method of modeling multiple-utility function. The subsequent issue is how to apply it in practice. This question leads to simple guidelines for defining a utility function.

5.5 Guidelines of Modeling Utility Functions

Based on the clustering parameters of six cloud market segments, as shown in Table 5—1, the type of business application can be estimated, which is mapping to each corresponding cloud market segment (See Figure. 5-3). If an analyst has the real cloud operational dataset, this step will become much easier. The essential issue is how to define the utility function for different customers' business applications? The basic guidelines can be summarized as follows:

1. If the business customers host a web site or run e-commerce applications, such as online checkout, one of the major value propositions for a customer to purchase more VMs is to reduce the queuing time. The process of reducing queueing time has been demonstrated. The effective model to describe the cloud customers' utility value is the power function for SME. However, the parameter of the exponent has to be negative to reflect the diminishing of return for the marginal utility. Figure.5-10 illustrated the value proposition for an e-commerce type of business application
2. When the exponent of the power function is equal to one, the power function becomes linear. To reflect the diminishing of return for the marginal utility, the coefficient value

of the linear variable is negative. The primary driver behind adopting a linear function for the VDI application is to increase storage performance and VDI scalability. According to [280][319], there are 19 performance metrics, such as "Copy Read Hits," "Disk Time," "Pool Paged Bytes," "Network Interface Bandwidth," etc. Different performance metrics might change the utility function parameters. It is dependent on CSP's targeted customers. The best practice is to set up an initial model and then have a fine-tuning of the model based on the real operation dataset.

3. If the exponent of the power function is set to zero, the function becomes a constant within the particular range of VM. This function can describe a cluster of VMs to support the specified SLA (e.g., 5 nines) for mission-critical business applications, such as CRM database backup. The utility function becomes a discrete function because the utility will diminish to zero after a certain quantity of VM.
4. In comparison with the database backup, the DR application needs more VMs to mirror the production environment. From a utility function perspective, it means the number of VMs is more than the backup application. The above example assumes the maximum number of VM. However, the coefficient "θ" of the function is less than one to reflect the possibility of risky disaster that may occur.
5. Regarding a risk assessment of a customer's operational cost (e.g., CSP's offering price for cloud resources) and a possibility of workload interruption (e.g., performance), we can use the isoelastic (power) utility function to model the business customers' decision in term of acquiring the number of VMs. If the assumption is that the customers prefer to constant relative risk aversion (CRRA) for their dynamic content delivery, then the α value is between 0 and 1.
6. In contrast, if the customers prefer to take more risks for multiple interruptions of their computational process (such as MapReduce application) rather than pay a high price of cloud resources, the exponential utility function can be applied, and the value of α is less zero. Usually, the type of cloud application often requires massive computing power, and job priority is quite lower.

Throughout this chapter, we mainly use analytic approaches to model multiple utility functions based on the scenario of six market segments so that CSP can explore a more addressable cloud market share. We can also combine this approach with a statistical approach to define the cloud customers' utility functions if we can access a live operation dataset from a CSP. In comparison

with the statistical method, many assumptions of our model can be further consolidated. The advantages of combining various methods would provide a much balance view of cloud customers' preference in terms of marginal VM demand because different cloud applications would have different utility values. Therefore, a combination of the statistical and analytic methods enables a CSP to know more about how much the customers are willing to pay for a particular type of cloud resource (e.g., VM instance).

The idea of the modeling cloud utility function based on the segmented market is to measure cloud business customers' preferences and tastes in terms of less or more VM resources to be purchased. In this study, the unit of subjective metrics can be interpreted as the cloud customer's revenue or profit contribution. Practically, many factors may impact the business customers' revenue and profits, such as end-user' experiences, response time, latency, throughput, availability, market environment, etc. Different measurements of CSM may result in a shape variation of a utility function. However, the above six utility functions cover some basic cloud business applications.

5.6 Summary

The issue of how to define the cloud customers' utility functions from the cloud customer's perspective is vital to any CSP because it would help the CSP to generate adequate cloud price models to maximize the profits for its cloud business. Based on the intensive literature review for this topic, one way to improve CSP's business revenue is to determine the cloud market segments first, and then to adopt the approach of customer value co-creation.

Throughout this chapter, we demonstrated how to construct different utility functions via value co-creation practice from a cloud customer perspective based on the segmented market. We also presented how to derive various utility functions from specified SLA, the response time of a checkout system, and degree of risk so that CSPs can build up various realistic customers' utility functions for different market segments.

To sum up, the modeling method can provide more realistic cloud customers utility functions, which is closely tied to the cloud customers' business applications. In comparison with previous modeling methods, our method is based on both market and customer value orientation. It is

flexible and practical for many cloud practitioners because all utility functions are measured by cloud customer revenue or profit values in terms of cloud resource consumption.

Chapter 6

Value-Based Cloud Price Modeling For Segmented Business to Business Market

Cloud price modeling is a significant challenge for many cloud computing practitioners and researchers in the field of cloud economics because the pricing of cloud services is often subjective and arbitrary. Many previous attempts mainly focused on a uniform market and used existing price models to explain the issue of revenue maximization for cloud service providers (CSPs) from an internal rationality perspective but paid less attention to the cloud market segmentation for cloud business customers from an external rationality perspective. This study considers both aspects of the value propositions. Based on the assumptions of the customers' utility values of different market segments, this research establishes a framework of value-based pricing strategy and demystifies the process of modeling and optimizing cloud prices for CSPs to maximize its profits. It shows how to create four cloud pricing models, namely: on-demand, bulk-selling, reserved, and bulk + reserved. It also illustrates how to identify the optimal price point of each model to maximize CSP's profit by genetic algorithm. This chapter demonstrates that bulk + reserved, on-demand, bulk-selling, and reserved can deliver a profit margin of 219%, 173%, 179%, and 213% for CSPs, respectively. Although bulk + reserved can achieve the highest profit margin, it does not mean that CSPs should adopt one model only because the cloud market is highly competitive. This chapter demonstrates a novel solution that CSPs can achieve the maximum profit with multiple pricing models that are offered to the segmented market simultaneously. This Chapter argues CSPs should capitalize on cloud pricing rather than price

---

This Chapter is derived from:

- **Caesar Wu** Rajkumar Buyya, and Ramamohanarao Kotagiri, Value-based Cloud Price Modeling for Segmented Business Market*, Journal of Future Generation Computer Systems (FGCS),* Volume 101, Pages: 502-523, ISSN: 0167-739X, Elsevier Press, Amsterdam, The Netherlands, December 2019.

to gain market competitive advantages. Thus, it provides state-of-the-art cloud pricing for segmented business to business market.

6.1 Introduction

Value-based cloud price modeling for different cloud market segments [170] is vital to all Cloud Service Providers (CSPs) as it will not only impact on CSP's profitability but also determine their business sustainability [10]. The goal of this study is to develop a comprehensive process framework of value-based price modeling that enables CSP to gain more cloud B2B market share for its profit maximization. Many previous studies can be considered as either cost-based or cost-plus models [75], which they were dependent on an assumption of a uniform market and paid less attention to the segmented market that carries heterogeneous values of customers. Furthermore, their processes of modeling mainly explained how to leverage two or three existing models (e.g., on-demand, reserved and spot instance) for CSP to maximize its revenue, which was subjective to a cloud capacity constraint that is equivalent to a cost. Subsequently, those works can be categorized as "pricing for the internal rationality."

The term of "Rational" means a decision is made according to reason or logic. In economics, people are assumed to be rational because they will systematically and purposefully do the best they can do achieve their purposes, given the available choices [320]. "Internal rationality" implies that a decision-maker focuses on internal justification; for instance, a cloud price is determined by a capital cost. In contrast, "external rationality" suggests that a decision should be made by an explanation of external factors, in which pricing is dependent on customer willingness to pay. In economics, it is essential that the pricing model is built upon the assumption that the individual is rational because people can be irrational.

The questions of how to create a cloud price model based on the business customers' value proposition and how to target the segmented market, especially, business to business (B2B) market, have remained either unanswered or incomplete. To overcome this gap, this chapter develops cloud price models that include both external and internal rationalities. For the external rationality, this model will include two essential external factors, namely, cloud customers' utility values and B2B market segments. For the internal rationality, this model takes consideration of CSP's cloud infrastructure cost. Based on the result of new price modeling, a genetic algorithm

(GA) will be applied to identify the optimal price point of each price model for CSP's profit maximization. One of the useful properties of GA is that it can solve a complex profit equation for intertwined variables without knowing the details of sub-functions. It is also convenient to upgrade the optimal price point of each model so that the process of price modeling can cope with the decision variation of cloud business strategy.

To demonstrate the complete process of value-based cloud price modeling, this chapter exhibits and analyze different models, namely cost-plus, on-demand, bulk-selling, reserved (two-part tariff), and reserved plus bulk for profit margin comparison. The cost-plus pricing models are also known as cost-based pricing, which is often prevalent [10] because "they carry an aura of financial prudence… to yield a fair return on overall costs (or resources), fully and fairly allocated." However, these models fail to capture the heterogeneous values of cloud business customers. In contrast, this chapter proposed four value-based models that can reflect the value proposition of both cloud customers and CSPs. Those models can be considered as "value co-creation" [271] [272] because CSPs are seeking a partnership with their cloud customers in the cloud market value chain. It shows these models allow CSPs not only to satisfy customers' needs but also to achieve a better profit margin in comparison with the cost-based model. Overall, this chapter provides a process solution that can measure both CSPs' profit and customers' utility under a single currency, which is customers' business revenue contribution. This revenue contribution can be defined by different cloud customer service metrics (e.g., increase sales, customer retention, investment efficiency, maintain a specified SLA, reduce checkout queueing time, etc.). To better illustrate the entire process modeling, this chapter exploits the following business scenario to explain the details.

6.1.1 Background

Assume a group of decision-makers of a hosting firm decide to expand its hosting business into the cloud B2B market. It implies that the firm wants to become a new CSP to compete with other existing CSPs (either global or local CSPs). If the initial investment budget (both capital and operation expenditure) and business goals (targeted revenue, profit, and market) have been approximately identified, the decision-makers want to know how to achieve the business goals. There are two fundamental questions must be clarified: "How does the firm form the right pricing strategy for the identified business goal?" and "how does it decide the adequate cloud price models along with optimal price points, sales volumes, and unit cost to achieve the maximum

profit?" These questions will help the CSP to divert its limited resources (investment budget and expertise) to serve its targeted customers better so that it can maintain its cloud business profitability and sustainability. There are many possible pricing strategies to reach the business goal, namely cost-based, market-based, and value-based pricing. As Hinterhuber [36] indicated, both cost-based (37%) and market-based pricing (44%) are much popular than value-based pricing (17%) shown in Figure 2—5. T. Nagle [10] observed that historically, cost-based pricing is the most common pricing strategy in most industries because "in theory, it is a simple guide to profitability; in practice, it is a blueprint for mediocre financial performance." Unfortunately, the issue with the cost-based pricing strategy is when there is strong market demand, the average unit cost will decline, and the price reduction should follow because the profit margin is determined by the unit cost (e.g., 30% ~100%). Conversely, when the market demand becomes weak, the average unit cost will go up, and the price should be raised. It contradicts a sensible pricing strategy in terms of market response.

The alternative way of cost-based pricing is either market-based (or competition-based) or value-based (customer-driven) pricing. Market-based pricing is to set a cloud service price based on the current competition condition of supply and demand. However, competition-based pricing could mislead CSPs to see market-based pricing as a zero-sum game, which what the customers' gain is the CSP's loss [201]. They could believe they do not influence price because market-based pricing is a competitive behavior of the market. In contrast, value-based pricing can offer customer needs and create real value to satisfy those needs of business customers because it can accurately quantify the customers' utility values.

Nonetheless, the definition of value-based pricing can be subject to a wide range of interpretations. It is dependent on the context of the content. The term is often defined as a pricing process for an individual's preference (ordinal utility [222]) that aims to the B2C market but, this chapter of value-based pricing focuses on the marginal value (cardinal utility) that aims to the B2B market. It implies the process of capturing a portion of CSP's economic impact on a target cloud customer's business [201]. In other words, a CSP is to develop and deliver the cloud service values for the cloud customer's business success and then seek a reward for the distributed services in the B2B market.

In general, the B2B market emphasizes the entire value chain and partnership development. The purchasing decision is not made by single or few individuals, but by more than dozens of

stakeholders for the cloud business values that CSP can offer. Therefore, value-based pricing is one of the effective pricing strategies for the cloud B2B market. The cloud services can influence the customer's business in terms of increasing their profit margin (cardinal utility), which is to achieve higher business revenue and lower the operation cost.

Overall, the processing framework of value-based pricing strategy includes 1) identifying target customers and workload patterns that are related to each cloud market segment, 2) quantifying cloud customer utility functions that are associated with service values and cloud service metrics, 3) establishing various cloud pricing models based on the specified customer's utility functions, 4) identifying the optimal price points for CSP to achieve the total maximization profit from all market segments. Figure 6—1 presents this processing framework of price modeling of all elements.

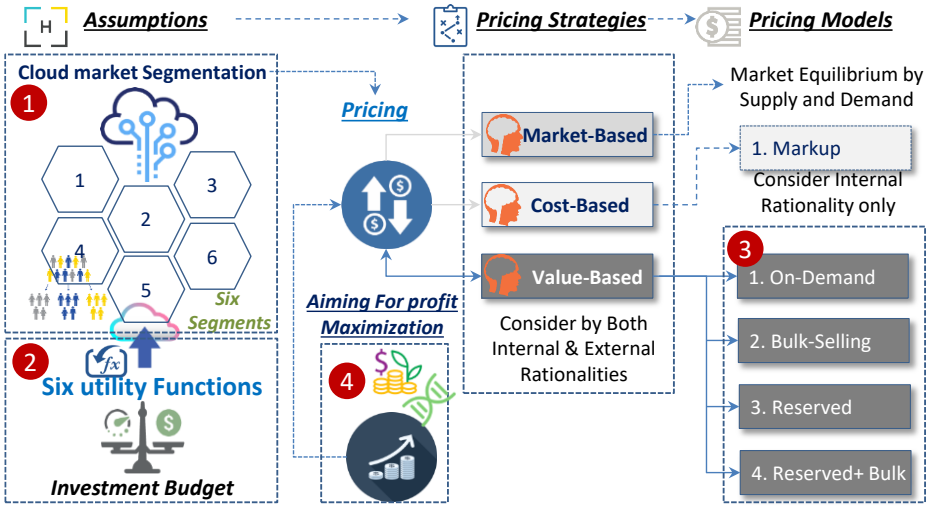


Figure 6—1 The Scope of the Problem

Due to the details that have been already included in chapters 4 and 5, this chapter will only focus on developing cloud pricing models (element 3) and determining the optimal price points (element 4). However, this chapter will include brief information on the cloud market segmentation (element 1, shown in chapter 4) and customer utility functions (element 2, shown in chapter 5).

6.1.2 Cloud Market Segmentation

The purpose of cloud market segmentation is to gather cloud customers' usage patterns so that a CSP can work out a good pricing strategy to anticipate a rapidly changing cloud market with

enough resources to serve its targeted business customers well for its profit maximization. In fact, Yankelovich [169] articulated the detail objects of market segmentation: align with the company's strategy, specify where the revenue and profit come from, articulate customer's business values, Focus on actual business behaviors, make sense to the firm's executive team and the board, and to be flexible to quickly accommodate or anticipate market changes. Based on these market segmentation criteria, chapter 4 has developed a novel solution that allows CSP to identify the cloud B2B market segment quickly. The solution is a combination of hierarchical clustering (HC) with time-serial (TS) methods based on two datasets, which one is downloaded from Google public dataset [178] and the other is extracted from a local hosting firm for its hosting services. From Google's dataset, we can develop six potential cloud market segments based on the number of parameters of cloud customers' usage patterns, such as job priority, number of cores, memory size, and AMD's virtualization workload guidelines [276]. This number of cloud market segments is within the range of McDonald's suggestion [172], in which the suggested number of the market segment is between 5 and 10.

The results of cloud market segmentation are shown in both Figure.5—2 and Table 5—1. Once the cloud market segments have been quantified, the next issue is how to develop the cloud customers' utility functions for the defined cloud market segments (Table 6—1)

Table 6—1 CLOUD CUSTOMERS UTILITY FUNCTIONS AND MARKET SEGMENTS [19]

Segment	Seg 1	Seg 2	Seg 3	Seg 4	Seg 5	Seg 6	Total
Average Job Priority [20]	1	0	2	0	3	3	
Average number of Cores	2	23	1	1	13	3	
Average number of Memory	7	6	6	3	102	86	
Perentage	30.1%	23.0%	10.0%	26.3%	9.1%	1.4%	100%
Predicted Sales Vol [21]	269	205	90	235	81	13	893
Estimated Possible Workload [22]	Static or Dynamic	Static or Dynamic	Static	HA	HA	Backend [23]	
Example of Apps	Web Hosting Server & Online checkout	Dynamic Content Delivery	Virtualized Desktop Infrastructure	Database Server	Disaster Recovery & BI	Backup, Logfile process	

[19] HA = High Variability, DR= Disaster Recovery, VDI = Virtual Desktop Infrastructure
[20] In this case, "job priority" carriers more weight for the decision of cloud workload pattern[233]
[21] Sales Volume is estimated by time serial (TS) predication without consideration of probability, which will be done in separated research work.
[22] The possible workload estimation is based on the recommendation of AMD's paper and cloud design patterns [264][233] [276]
[23] Backend type of workload patterns might also include business intelligent (BI) or log data analysis[276]

6.1.3 Modeling Cloud Customers Utility Functions

The goal of modeling cloud customers' utility function [261] is to quantify the cloud customers' subjective preferences (or utility values) that are subject to the cloud resources acquired. Practically, the subjective preferences concern the customers' business applications for the revenue contribution, which can be measured by the operational metrics [285].

The meaning of utility is quite ambiguous because it consists of different connotations. Historically, the implication of utility was derived from utilitarianism. It means a subjective experience and satisfaction. It is known as the utilitarian tradition. Later, this term has been extended to the contractarian tradition, which emphasized social welfare [321]. As a result, the contemporary meaning of utility has three connotations:

1) The economic utility refers to subjective satisfaction and happiness. "It is an alternative way to describe preference and optimization [320]. The utility value in this context is measured by different preferences under information uncertainty in terms of risks and wealth.
2) Another implication of utility is an essential infrastructure service for the public. Sometimes, it is also called as "public utility," such as water, electricity, and telephone service that are supported by some incumbent providers. It is associated with the term of social welfare
3) "Utility" also refers to the utilization rate. It is measured by a percentage value between 0 and 1. For example, the utility of a network means its utilization rate. It is a concept of efficiency. It is different from the economic connotation of utility that is measured by preferences.

However, there are many previous works that assume both economic utility and utilization rates are the same. The utilization rate can be included in a cloud service metric, but it is not the same as the utility value in an economic sense. Economically, a business customer's utility represents the amount of business revenue or profit that is contributed by the number of VMs (e.g., wealth) that can support business application workloads. For example, the utility of mission-critical or high availability (HA) application workloads (e.g., sensitive data, liability if breaded or deleted) will be totally different from the utility of the backend type of workloads, such as MapReduce [285]. The end-users will pay a different price for different applications. For example, the MapReduce workload can be interpretable. The question is, how we can use a single currency to reflect various utility values and align with CSP's profitability? To solve this issue is to unify all

customers' utility values and CSP's profit into a measurement unit, which is cloud customers' business revenue or profit. This is also known as value co-creation. The benefits of value co-creation are that CSPs can reduce investment risk and maintain cloud customer loyalty [303] and uphold CSP's profitability and business sustainability. The modeling process of quantifying customers' utilities is to establish a relationship between the customer's business profit contribution (a dependent variable) and the number of VMs (independent variable) required.

Based on different characteristics [288] of the cloud business applications, we organize these utility functions into three categories:

- Utility functions (Segment 4 and 5) are defined by High Availability (HA) characteristics [220] [263] [291]
- Utility functions (Segment 1 and 3) are determined by response time characteristics [286]
- Utility functions (Segment 2 and 6) are identified by risk characteristics (risk-averse, risk-seeking, and risk-neutral [285])

The process of how to quantify these utility functions has been presented in chapter 5. Table 5—4 highlights the result of six utility functions (assumptions of utility functions presented in Section 6.3.2.1). Now, the subsequent questions are how we can generate various price models for a CSP to capture more cloud market share and how to identify the optimal price point of each model for profit maximization? These problems will be solved in this chapter.

6.1.4 Problem Definition and Solution

By microeconomics [7], the problem of cloud business profit can be formalized as the total business revenue minus the total cost (Equation 6-1 and 6-3). The total business revenue is dependent on a sales price, an average unit cost, and sales quantity (or market demand). Intricately, a sales quantity is a function of a price, and a price is an inverse function of the sales quantity. Mathematically, it can be presented in the interdependent relationship in Equation 6-2.

$$\pi[p] = R[p] - C[Q(p)]\,, \qquad C[Q(p)] = c_u[Q(p)] * Q(p) \tag{6 -1}$$

$$R[p] = p * Q[p];\ p = Q^{-1}[p] \tag{6-2}$$

$$\pi[p] = Q[p] * \left(p - c_u\left[Q[p]\right]\right) \tag{6-3}$$

where $\pi[p]$ is a cloud business profit for a CSP, $R[p]$ is a cloud revenue, $C\ [Q(p)]$ is the total cost, p is a unit price and $c_u[Q(p)]$ is the average unit or marginal cost which is also a function of the total sales quantity $Q(p)$.

The issue is how to achieve the maximum profit by identifying the optimal price point (Equation 6-4). While the equation appears evident and straightforward, it is difficult to find a clear solution because of both functions $Q(p)$ and $p = Q^{-1}(p)$ are generally unknown

$$p^* = \underset{p}{\text{argmax}}\ \ \pi[p] \tag{6-4}$$

The primary challenge is that the relationship of $p = Q^{-1}(p)$, $c_u[Q(p)]$, and $Q(p)$ is intertwined. Moreover, these equations will become progressively more complex if various pricing models are introduced.

Previous works solve the problem by excluding the cost component from a profit equation 6—1 [55] or by making some restricted assumptions [56] [252], or by assuming a uniform market that is derived from α-fair utility [80]. Others assume a price is a simple linear equation based on the AWS' historical data within a coefficient band [61]. Still, others concentrate on cloud spot instances [56]. Although their works have made excellent progress in the context of cloud price modeling for the B2C market, many critical aspects of modeling remain unanswered. This chapter provides the solution for the issues (elements) of 3 and 4 shown in Figure.6 -1, and the solutions to these issues encapsulate the complete process of value-based pricing strategy.

6.1.5 Contributions

By providing the above solutions, this chapter has made the following contributions:

- To the best of my knowledge, it is the first time to create various cloud price models based on market segmentation theory and the number of utility functions that are defined by cloud customer business impacts.
- This chapter has clearly illustrated how to establish four value-based price models according to the defined firm's business strategy
- By leveraging the existing retail pricing experiences (such as Costco), this study develops a bulk-selling+ reserved model for a CSP to achieve the highest profit margin.

- This chapter also illustrates the relationship between bulk-selling and bundle services. By developing various cloud pricing models, CSPs have more pricing options to capture more profit cross various market segments.
- This chapter demonstrates how to apply GA to identify the optimal price point for each price model.
- The price models are dependent on both internal (CSP's cloud infrastructure costs) and external (cloud market segments and customer utilities) rationality.
- This chapter presents a novel and practical solution with the framework of price modeling so that many practitioners can plug in their datasets and then build their price models based on the defined company's business strategy.
- Most importantly, this chapter shows how to calculate the total revenue and profit based on different pricing models that are offered to various customers spontaneously.

6.1.6 Chapter Organization

The rest of the chapter is organized as follows: Section 6.2 provides a brief overview of the literature regarding cloud price modeling. Section 6.3 formalizes four value-based pricing models with various assumptions and constraints. Section 6.4 presents concise information about genetic algorithms (GA) and how to determine the GA parameters for the experiments. Section 6.5 shows the experimental results. Section 6.6 offers a detailed analysis and discussion of cloud price modeling and optimizing. Section 6.7 provides a summary of this chapter.

6.2 Related Work

In light of the value theory or axiology [100], this chapter approximately classifies the cloud price models into three basic pricing categories, namely value-based pricing, market-based pricing, and cost-based pricing. The value-based pricing is often considered as a subjective view of the cloud pricing from a demand-side because it concerns the measurement of customers' subjective experience or preferences. The cost-based pricing model is regarded as an objective view of cloud pricing from a supply-side because it is built on the physical quantity of unit cost. The market-based pricing is an interactive view of both value-based and cost-based pricing for the equilibrium of supply and demand in the marketplace. According to this classification, it is

easy to identify most of cloud pricing models can be considered as either market-based or cost-based models [293].

For example, Macias et al. [250]used a genetic algorithm method determining a cloud price. Their model can be classified as market-based pricing. The study aims to offer a solution to a competitive price for the negotiation of the services market. However, they recognized their work has some limitations. They believe "it is difficult to establish a profitable pricing function." This work shows how to overcome this limitation and bridge this gap in the later sections. Although Macias et al. [250] made some progress in term of modeling the cloud utility function for SLA metrics, one of the critical issues has remained unsolved, which is how to include the demand side's utilities or cloud customers' business values for CSPs to generate various cloud pricing models and to achieve a partnership with cloud customers [304].

Kilcioglu et al. [251]present a calibrated benchmark model for cloud pricing based on empirical data. Their model can also be categorized as one of the market-based pricing models. Kilcioglu et al. [251] explained the market trends of the cloud price and higher profit margin of AWS based on the quality competition assumptions under both monopoly and duopoly market environment. The chapter showed the utility function of the cloud customer consists of three elements, subjective values, delay sensitivity, and service quality.

It was the first time that the demand side's utility function had been defined as a function of both subjective values and objective costs [251]. The paper made a good contribution to the theoretical modeling of price-quality competition in both a monopoly and duopoly competition market. Nevertheless, many problems are still unanswered, such as the determination of subjective values of utility functions for the cloud B2B market.

Aazam et al. [252] established a resource-based price model by cloud customer's historical pricing record for digital media stream workload across an inter-cloud environment (via cloud brokers). Although the authors made a great effort to build a model equation for inter-cloud pricing, many critical values of the equations are restricted to some particular cases, such as a data stream type of workload. However, the authors provided a framework for modeling and analyzing AWS on-demand and reserved instance pricing based on historical observation.

Yeo et al. [253] argued that automatic metered pricing model for a utility computing service (computing service as a commodity) could achieve a better revenue result in comparison with

fixed pricing, fixed-time, and Libra [255] plus dollars $ [256] (a pricing model based on the users' requirements). The paper presented a compelling pricing model for self-justification, but more experiments are required, as the authors indicated. Xu et al. [80] presented a similar idea and developed various pricing models (such as the 1st order discrimination, resource throttling, energy (or cost) saving and SLA charge) to maximize CSP's revenue that is subjective to CSPs cloud infrastructure capacity and customers' surplus value. The authors argued that the usage price depends on the utility level distribution and the elasticity parameters α on the base of their theoretical proof for Theorem 1 (see Equation 6-5 by leveraging KKT condition[322]) Although their utility connotation was referred to economic utility, this α was derived from the α-fair network utility rather than a customer's preference. They concluded that pricing discrimination had no effect on CSP's revenue maximization

$$p_v = \frac{\lambda}{1-\alpha} \tag{6-5}$$

This conclusion contradicts Claycamp and Massy's [170] theory of market segmentation and McDonald's, the practical solution of market segmentation [172]. There are some gaps in terms of Xu's work.

1. The economic sense of isoelastic utility function has different meanings of α-fair network utility because the earlier one is to measure a subjective experience and the latter one means the efficiency of utilization rate.
2. The optimal price: p_v is dependent on the variable of Lagrange multiplier λ, which is not defined.
3. As a result, the α-fair parameter is not inversely equal to price elasticity of demand of an isoelastic utility function.

$$E_d = \frac{\partial Q(\cdot)/Q(\cdot)}{\partial p/p} \tag{6-6}$$

where $Q(\cdot)$ is the quantity of the demand good? The α-fair utility means a priority of time scheduling while the α of isoelastic utility means the degree of risk. As Xu et al. [80] indicated, their work was an extension of Hande et al. [177] study of pricing access networks with capacity constraints for revenue maximization.

Before Xu's paper, Joe-Wong and Sen [113] had also proposed a similar solution to a cloud pricing strategy that is subjective to the cloud capacity. The root of their pricing strategy was also

derived from the access networks. The purpose of their research work was to develop an analytic framework to balance the fairness (welfare concept) of resource priority and CSP's revenue maximization by various pricing models. Although there were some differences between them (e.g., Xu's work included a probability of utility level distribution, and Joe-Wong discussed fairness), both studies assumed there was a uniform market and corresponded to a α-fair utility function. All studies relied on the Lagrange multiplier or Karush -Kuhn-Tucker (KKT) conditions to identify the optimal price point, which is subjective to the specified limited capacity. Ultimately, these works used the analytic tool to prove there is an optimal price point.

Recently, Shahrad et al. [118] proposed an incentive pricing solution by balancing limited cloud capacity and demand peak time. Shahrad's core idea is to leverage the cloud price as an incentive to regulate the usage behavior of cloud business customers, which means they would allocate cloud resources by themselves according to CSP's price variation. It is a self-regulate idea to eliminate its own demand during a peak time and fill its workloads during a valley time. The customers' utility function is the same as α-fair one.

All these studies assumed one type of utility function that is α-fair network utility for cloud customers. All papers assumed that economic utility and the utilization rate of a network are equivalent. In order to achieve maximum profit, the objective function must be differentiable. In contrast to the α-fair network utility function, Chen et al. [69] proposed a utility function that is driven by the cloud customer's satisfaction in term of price and response time shown as follows:

$$U(p,t) = U_0 - \alpha p - \beta t \tag{6-7}$$

where U_0 is the maximum utility value and both α and β and constant coefficients. Price p and response time t are two independent variables to reflect different levels of utility value or customer satisfaction. If both p and t is equal to zero, it means the customer has maximum utility value. This is a linear utility function. The response time can be represented in price (or a cost) p because if CSP adds more resources, e.g., VMs for workload process, the response time t can be reduced. In addition to this issue, the paper did not give the optimal price point between CSP's profit margin and cloud customers' surplus values (customer preferences).

In comparison with creating a new pricing model, other research works [55] [63] [61] intended to extend the current cloud pricing models offered by different CSPs for profit maximization. Xu et al. [56] combined both reserved and spot instance prices that allow a CSP to maximize its

revenue and profit through a dynamic cloud pricing model. The work was derived from empirical observation of the historical price of Amazon Web Services (AWS). The paper made contributions to an alternative pricing model for a spot pricing scheme. Following a similar line of reasoning, Alzhouri F. and Agarwal A. [275] constructed a theoretical or dynamic pricing scheme for CSPs to maximize their revenue via a solution of a dynamic programming approach. The potential issue of their revenue maximization without consideration of the average unit or marginal cost would become economically unsustainable. Toosi et al. [55] consider all three types of pricing models, namely on-demand, upfront reserved, and a spot for CSP's profit maximization, but the unit cost of cloud resource remains untouched [254]. Brynjolfsson et al. [305] argued that this kind of cloud pricing could be "overly simplistic … blinding us to the real opportunities and challenges of cloud computing."

On the other hand, Ben-Yehuda et al. [61] suggested the price of AWS spot instance is not driven by some market mechanism or an auction approach rather than it is randomly generated from a tight price range that has a dynamic hidden reserved price mechanism. This indicates that the price mechanism of AWS spot instance (2 minutes notification) is similar to Google's preemptible VM instance (80% discount but terminated after 24 hour execution time with 30-second notification), and Azure low-priority VM or eviction instance (with 60% (for Windows)-80% (other OS) discount, excluding B-Series VMs, 30-second notification), which means the spot price has either implicit or explicit bottom-line. The problem with these instances (or VMs) is that both service availability and capacity cannot be guaranteed. Moreover, many new service features are excluded. Perhaps, MOZ's [73] business experiences [24] in 26-Sep-2011 provided a good lesson for many cloud business customers. The incident suggests that the spot instance could be an unreliable cloud resource for some mission-critical applications. Overall, the previous works can be summarized in Table 6—2 in terms of main contributions, advantages, and gaps

[24] MOZ reserved bid for AWS spot instance was $2/per instance for more than 3 years

Table 6—2 SUMMARY OF SOME PREVIOUS WORKS

Category of Pricing Models	Main Contribution	Advantages	Potential Gaps
Toosi et al.'s Heuristic Algorithm of pricing model [55] (2014)	It combined three different pricing models for profit maximization for CSP profit maximization	Consider all available pricing models at that time	Excluded cost component.
Ben-Yehuda, et al. Statistical regression (2013)[61]	It provided a rough estimation of the AWS pricing model for spot instance	Proposed alternative solution for the pricing model	Observation of historical records. Lack of rationality
Hande et al. α-fair utility model [177] (2010)	It introduced one of the utility functions for the pricing model	Highlight price elasticity and utility function	Ambiguity definition of Utility and pricing Elasticity
Xu, Hong, and Baochun Li [80] α-fair utility model (2013)	It introduced the probability density function for cloud customer demands	Show KKT Proof	Contradict to Market segmentation theory
Joe-Wong et al. α-fair utility model [113] (2012)	It offered a mathematics framework for cloud pricing	Introduced multiple pricing models for cloud pricing	The only proof of the optimal price without consideration of market
Shahrad et al. [118], Cobb-Douglas $\rightarrow$ α-fair utility model (2017)	It proposed a novel idea of increasing cloud data center capacity utilization rate while to maximize CSP' profit	Show Euler homogeneous proof	Utility function has to be differentiable
Chen et al.Customers' Satisfaction linear Utility model [69] (2011)	It introduced a linear utility function for cloud pricing	It included both price and SLA level into the utility function	Not clear in term of the optimal solution for CSP's profit maximization

As Kash, I A. and Key P. B. suggested [306], the spot instance price model has been attracted much attention in the academic world for cost-saving. Despite that, "the right answer remains unclear" [306]. One of the reasons is many price schemes are restricted to a particular case or application — for example, Jain et al. [88] suggested a value-based price model by leveraging the spot instance discount, but the model is only designed for batch workloads. In other words, different models could have different purposes with different functions. To visualize all pricing models with different purposes and functions, Table 6—3 highlights these differences among different models proposed by previous studies.

Table 6—3 DIFFERENT PRICING MODELS COMPARISON

Purposes with various Functions of Pricing model Comparison	Model Explain	Model Creation	Differentiable object function	Non-Differentiable	Market Segmentation	Max. Rev.	Including Cost element	Profit optimization	Optimizing Algorithm	Optimal price point
Macias et al. [250]	√			√		√			√	√
Kilcoglu et al.[251]	√	√	√	√			√	√	√	√
Aazam et al. [252]	√							√		
Yeo CS. et al.[253]		√		√		√	√	√	√	
Xu et al.[80]		√	√			√			√	
Hande et al[177]	√	√	√			√			√	√
Joe-Wong et al. **Error! Reference source not found.**	√	√	√			√	√	√	√	
Shahrad et al. [118]	√	√	√			√	√	√	√	√
Chen et al.[69]	√	√	√			√	√	√		
Toosi et al. [55]Xu et al. [56]	√					√			√	
Alzhouri et al. [275]	√		√			√			√	√
Ben-Yeuda et al. [61]	√			√			√		√	
Kash et al. [306]	√	√							√	
Jain N [88]		√		√					√	
This model	√	√	√	√	√	√	√	√	√	√

Although many researchers in this field have made excellent contributions to cloud economics, there are still many questions remaining unanswered: such as How to generate more price models for various cloud applications that can capture more business market share? How to practically identify the optimal price point for each model? How to translate multiple dimensions [306] of cloud service metrics (utility values) into a single currency between cloud customers and CSPs? How to address CSP's concerns about the cloud B2B market? How to create a value co-creation solution for both cloud customers and CSPs? How to determine the maximum profit with multiple pricing models in the segmented market? This chapter, together with chapter 4 and 5, provide a completed solution to these questions.

6.3 Cloud Price Modeling and Model Assumptions

6.3.1 Market Assumptions

According to the theory of the B2B market [273], the cloud B2B market is a relational business market because it emphasizes building a mutual value-generation relationship or a partnership with business customers. It requires long-term relationship development. In contrast, business to consumer (B2C) market mainly is focusing on the final transaction between the firm and end-user [273]. From this perspective, this chapter will first consider the cloud price models based on the assumption of a monopoly market [7] because the B2B market is much challenging for other market competitors to access the existing market [274]. Furthermore, many innovative characteristics of cloud services are often new to the current market. Chapter 3 provides a solution for establishing a cloud price model for innovative cloud service characteristics, which have no existing market. This is not a prohibitive assumption.

In addition to the monopoly assumption, this chapter also assumes the cloud market is not a single and integrated market rather than the segmented market because cloud customers have heterogeneous cloud applications. This assumption allows CSP to capture more cloud market share. Cloud market segmentation is to group personalized prices for heterogeneous demands so that the CSP can achieve the best profit margin within its resource capacity [170]. One of the typical examples of market segmentation in the service industry is the airline ticket price. The airline companies often classify their market into three or four segments, that is the 1st class, business class, economy, and cheap flights with different airfare prices and service conditions. Similarly, the cloud market can be grouped into different segments based on the different characteristics of cloud services.

6.3.2 Assumptions of Quantifying VM Resources

Following the segmentation result and the virtual server workload guidelines [276] [277], this work can approximately estimate the workload pattern of each cloud market segment, such as web hosting, high availability, backend data process, disaster recovery, content delivery, etc. [309] as shown in Figure 6—2. The number of VM quantity for one type of VM is equal to q , such as Amazon Web Service's instance of m4 extra-large or Google's ni-highmem-16. This quantity may vary from customer to customer. It is dependent on a type of VM instance and cloud business application. Based on consideration of all these factors, this chapter sets this maximum number is equal to 12 ($q_m = 12$) because we mainly focus on the small and medium enterprise (SME)

customers so that this maximum quantity is justified for a typical SME's application. This number can be either increased or reduced. It is just a matter of a scale.

6.3.2.1 Pricing Models Assumptions

6.3.2.1.1 Cost Assumptions

Along with the cloud market assumptions, this chapter also assumes the initial investment budget or Capex for one type of VM. The Capex and Opex ratio is 1:4. This ratio is based on local empirical data. The Capex is estimated by the latest average price of server hardware that is offered by major vendors, such as HP (HP Enterprise DL380, 2RU), Dell (PowerEdge R730), IBM (8203-E4A5634), and Cisco server (UCS M5). This study also includes some cloud data center installation costs [75] , which are shown in Table 6—4.

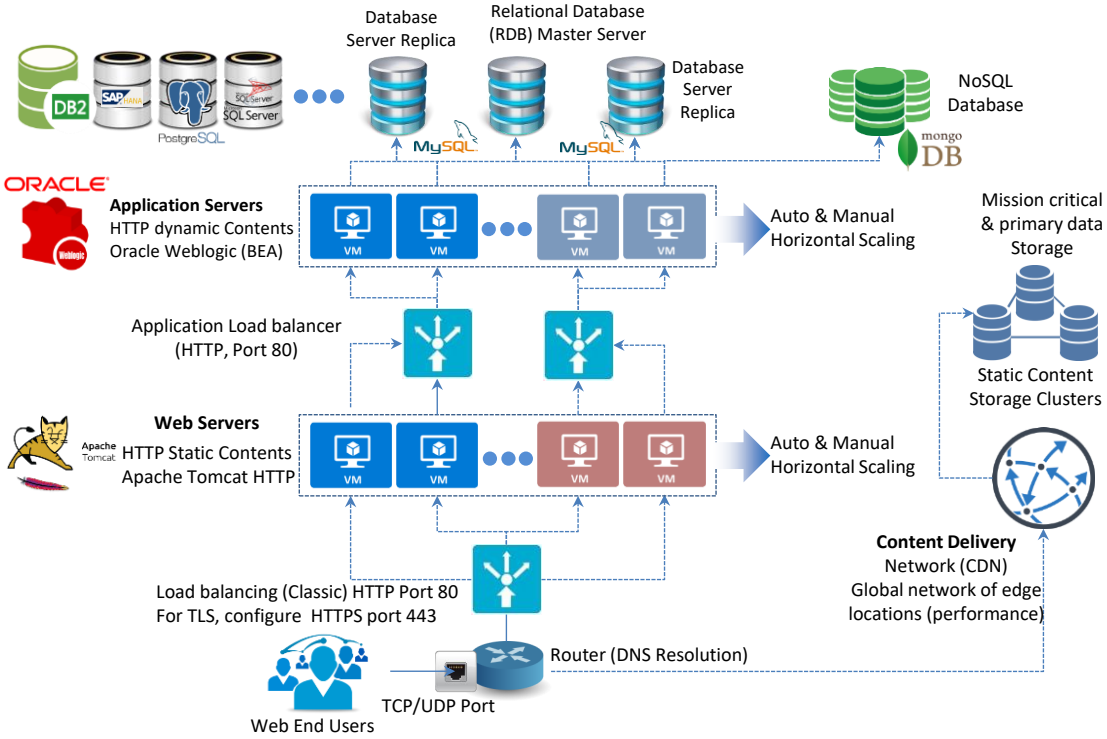


Figure 6—2 A Typical Architecture of Web Application Hosting

Table 6—4 Cost Assumptions

Capex/per hour	Opex /per hour	Capex & Opex Ration	Number of Physical Servers	Configuration	Number of VMs Capacity
$325	$1,300	1:4	500 - 600	8 or16 cores/per server	9,000 – 12,000

Note:
- Assumptions of investment Budget or Capex C = $3 million
- The number of physical servers ≈500 – 600
- The configuration per server is either 8 – 16 cors/ per server

6.3.2.1.2 Utility Function Assumptions

Table 6—1 provided the cloud market segment and predicted sales quantity, but they do not tell the cloud customer utility function of each market segment. To optimize the cloud pricing models, the cloud customer utility function for each cloud market segment should be defined. According to Krugman and Wells [224], different individuals would have different utility functions because different people would have different tastes and preferences. The essence of a utility function is to describe how people consume various quantities of goods in terms of their subjective preference and tastes (or utility) in a less or more rational way.

If the assumption is for CSPs to target the SME customers and focus on building mutual value generation, the modeling process is to define how their cloud resource (VM) can create SME's business profits. The effective modeling is that CSP should translate various cloud service metrics (Response time, SLA, end-users retention, and leverage investment) into a single currency (business profit), which is also in line with CSP's business value proposition. As a result, the cloud customer utility function is defined by the business customers' profit gain (surplus value) and cloud resources. The following equations are described their relation:

$$B_i = K_i\left(\sum_{q=1}^{q_m} u_i[q]\right), \quad i = 1 \cdots S \tag{6-8}$$

$$K_i = B_i/\left(\sum_{q=1}^{q_m} u_i[q]\right), \quad i = 1 \cdots S \tag{6-9}$$

where B_i is a yearly data. It represents customer business revenue or profit. If the Australian Bureau of Statistics (ABS) data for small businesses [237] is applied, a certain profit range for the targeted SME can be specified. For example, if the average net profit is approximately $39,000 and $90,000, the values of B_i across all segments can be calculated, as shown in Table 6—5.

$U_i[q]$ is a customer's utility function for the "*i*" market segment. "q" is the quantity that the customer will purchase. K_i is the scaling coefficient that reflects the utility level that is associated with a cloud customer's business profit. Further details will be illustrated in Figure 6—3.

Table 6—5 Cloud Customer Surplus Values in Six Market Segments When $\boldsymbol{p}^*$=$1

Customer's Proft or Surplus $\boldsymbol{B_i}$	41,000	90,000	80,000	80,000	80,000	39,000	**Total**
Utility Functions $\boldsymbol{U_i(q)}$	$\boldsymbol{U_1(q)}$	$\boldsymbol{U_2(q)}$	$U_3(q)$	$U_4(q)$	$U_5(q)$	$U_6(q)$	
q = 1	$1.50	$0.00	$1.50	$1.50	$0.75	$0.01	
q = 2	$0.75	$0.21	$1.36	$1.50	$0.75	$0.03	
q = 3	$0.50	$.38	$1.22	$1.50	$0.75	$0.05	
q = 4	$0.38	$0.54	$1.09	$1.50	$0.75	$0.08	
q = 5	$0.30						
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
q = 11	$0.14	$1.39	$0.14	$0.00	$0.75	$1.07	
q_m=12	$0.13	$1.50	$0.00	$0.00	$0.75	$1.50	
Customers Market Demand D_i	269	205	90	235	81	13	893
Cloud Workload Patterns	Web Hosting Server	Content Delivery	Virtualized Desktop	HA	Disaster Recovery	Log processing	

6.3.2.1.3 Risk Assessments

Risk assessments refer to a utility function is defined by cloud customers' preference for different levels of satisfaction for their business profit gain in terms of their attitude towards risk to vary with the amount of VM resource. For example, according to the cloud customers' usage pattern, the 2nd market segment is for the cloud customers to deploy the web content. It is a network-oriented utility function. According to [80] [177] [279] [292], the iso-elastic utility function can describe the customers' utility in term of the cloud resources requirement (Equation 6-10 and Equation 6-11):

$$U_2[q] = K_2 u_2[q], \qquad u_2[q] = \frac{q^{1-\alpha} - 1}{1-\alpha} \quad \alpha \in (0,1) \tag{6-10}$$

$$K_2 = B_2 / \left(\sum_{q=1}^{q_m} u_i[q] \right) = B_2 / \left(\sum_{q=1}^{q_m} \frac{q^{1-\alpha} - 1}{1-\alpha} \right), \qquad \alpha \in (0,1) \tag{6-11}$$

where "q" is the number of VMs, and α is the constant coefficient, which is set to be 1/3 [80]. The coefficient α is also to measure the degree of relative risk aversion. In this case, the cloud customers' utility value is assumed to be dependent on the measurement of constant relative risk aversion (CRRA) [283] for content delivery applications workload. Along the same line of

reasoning, the customers' utility function in the 6th market segment can also be created as an exponential function [222].

$$U_6(q) = K_6 \frac{(1 - e^{-\alpha q})}{\alpha}, \qquad \alpha \neq 0 \tag{6-12}$$

Here, the assumption is that the customers of this segment become risk-taking because the application (e.g., MapReduce) workload can be interrupted. The reliability and capacity guarantee of a cloud resource is not a significant issue. Cost-saving becomes the main priority. Therefore, the coefficient α is negative.

6.3.2.1.4 Utility Functions Based on High Availability

The high availability (HA) business applications require the mission-critical cloud infrastructure. If the assumption of the downtime is less than 5 minutes / per annum, then the service level agreement (SLA) must be higher than five-9s (or 99.999%). Based on Markov Chain analysis [310], the number of VMs to delivery the HA cloud application can be specified. If the VM quantity is more than this specified number, the utility value will be diminished to zero. Moreover, all VMs have the same utility value because these VMs guarantee SLA delivery together. Consequently, the utility function of the 4th segment can be defined as follows:

$$U_4(q) = \begin{cases} K_4, & 1 \leq q \leq k \\ 0, & k < q \leq q_m \end{cases} \tag{6-13}$$

where k is the specified quantity of VM to guarantee cloud applications' SLA. q_m is the largest quantity that customers will purchase [284]. Similarly, the utility function of the 5th market segment can also be created. The difference between the 4th and 5th segments is the customers of the 5th segment might have its own existing cloud infrastructure. They only purchase specific cloud capacity if the price is below a specified threshold level θ in comparison with their own infrastructure costs.

$$U_5(q) = \theta K_5, \qquad 1 < q \leq q_m \tag{6-14}$$

6.3.2.1.5 Utility Functions Based on Queueing Time

In addition to the mission-critical workload applications, the utility function for the e-Commerce can also be modeled by a Markov Chain process. The basic idea of modeling the utility function for the 1st segment is to reduce queueing time [221] [264] [287]. The following equation can estimate the utility function.

$$U_1(q) = K_1 q^{-c}, \tag{6-15}$$

where c is a constant value, this study sets the "c" is equal to 1 in this case because of the pattern of e-Commerce (e.g., checkout). Alternatively, a linear function can be adopted as a solution to describe the customer utility function for the 3rd market segment or VDI. There are many VDI performance metrics of a hosting environment regarding users' experiences, such as the peak of Input/Output Per Second (IOPS), storage capacity, response time, Read/Write ratio, future growth, etc. If these metrics have been prefixed during the Proof of Concept (PoC) period before VDI rollout, the additional VM will only add Opex and a burden to the cloud customers. So, we can use a linear model [311] [282] to approximate the cloud customers' utility values.

$$U_3[q] = K_3(rq + q_m), \qquad r < 0 \tag{6-16}$$

where "r" is a constant, but it is negative, which reflects the diminishing return.

6.3.3 Finding Optimal Price Point for Profit Maximization

Once all utility functions have been established, the next step is to create different price models for CSPs to maximize their profits. The following example illustrates an overview of the process to identify the optimal price point of price model for CSP to maximize its profit (See in Figure. 6—3)

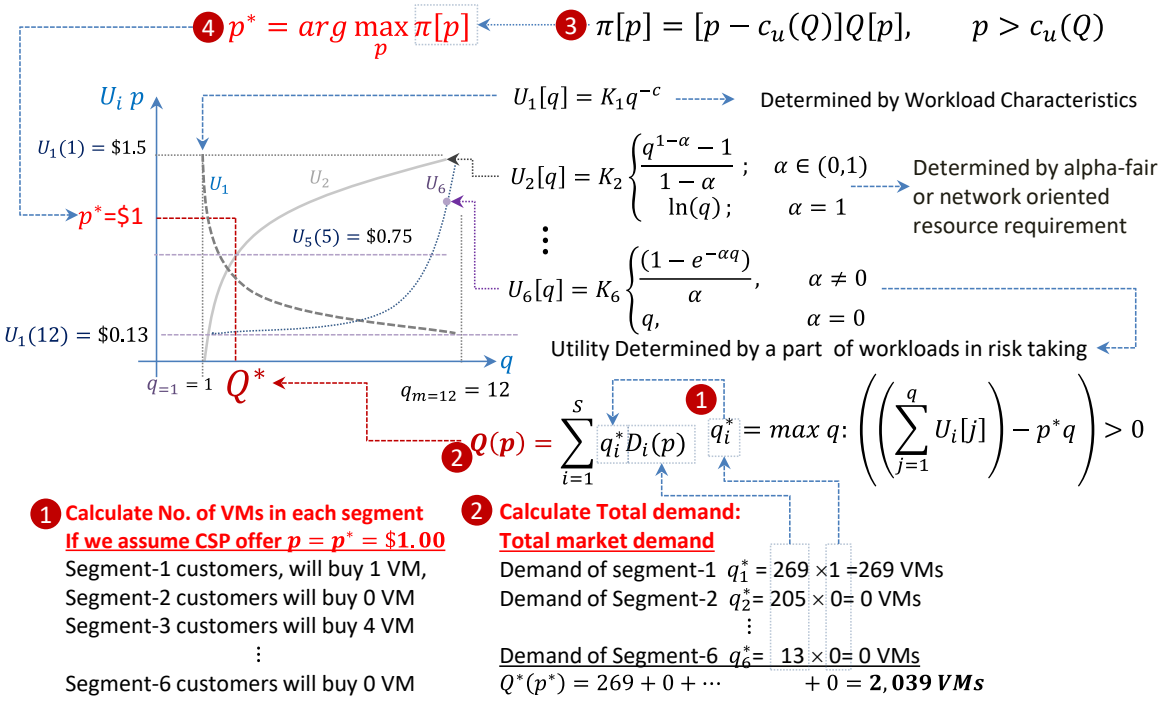


Figure 6—3 Overview of Optimizing Price When CSP Offer $\boldsymbol{p}^*$=\$1

Suppose a CSP offers \$1/per VM as its optimal price point (this price point is randomly selected. It could be the optimal price point for the CSP's profit maximization, but we don't know yet at this stage), we can calculate the cloud customer surplus values and a quantity of VM sales in each segment and the total market demand. Based on the defined utility function of the 1st segment, the cloud customers will purchase 1 VM, but not buy 2 VMs because 2 VMs would cost \$2, and the net surplus utility value of 2 VMs is only \$0.25 (\$1.5+\$0.75-\$2 =\$0.25), which is less than \$0.5 (\$1.5-\$1=\$0.5) for 1VM. In other words, if each customer of the 1st market segment buys 1 VMs and the total number of cloud customers is 269, then the total sales volume of VM is 269. Likewise, the customer of the 2nd market will not purchase any VM, but the 3rd segment will acquire 4 VMs. If we sum up all the VMs of all market segments, we can find the total volume of VM sales Q. As a result, we can calculate all the variables, including unit cost, profit margin, and total sales revenue. However, if this price point is not optimal, the question is how to identify the optimal price point for profit maximization across all market segments? Before answering this question, let us think about "are there different pricing models to achieve a better profit margin?" This question takes us to the topic of building various cloud pricing models in comparison with cost-based pricing.

6.3.4 Mark-up Pricing Model

As Hinterhuber indicated [36], the cost-based pricing is still prevalent in most industries, which is over 37% of pricing models. If the assumption of the mark-up price is 100% of the average unit cost or marginal cost, the expected profit margin would be 100% (Equation 6-17). The process of determining a price is very straightforward. On the flip side, this pricing model could be either overshot or undershot due to the pricing without external rationality.

$$p[Q] = mc[Q] + \frac{Q}{|\partial Q/\partial p|} \tag{6-17}$$

where $p[Q]$ is the price, $mc[Q]$ is the marginal cost, Q is the total demand quantity, $Q/|\partial Q/\partial p|$ is the markup price or profit margin. The price point is determined by the internal rationality or a CSP's desired profit margin or mark-up price $Q/|\partial Q/\partial p| = 100\%$.

6.3.5 On-demand Pricing Model

Alternatively, a CSP can build an "on-demand" price model [225] that is determined by both external and internal rationality. Many leading CSPs offer this price scheme. It is also known as Pay as You Go (PAYG). Usually, CSPs would charge at an hourly unit-based price. While both Google Cloud Platform (GCP) and Microsoft Azure use a sub-hour rate. Azure is 1/60th hour or per minute base, and GCP is a 1/6th hour or per 10 minutes base [258]. The sub-hour price should give cloud customers more flexibility and scalability to run various types of cloud workloads for "on-demand." In this chapter, the model adopts the hourly base unit. Based on the example of both Figure 6—4 and utility functions are shown in Table 6—5, the following equations can calculate the cloud customer surplus values (external rationality) 6-18.

$$q_i[p] = q_i : \max\left(\sum_{j=1}^{q} U_i[j]\right) - pq \geq 0, \qquad Q(p) = \sum_{i=1}^{S} q_i[p] D_i[p], \quad i = 1 \cdots, S \tag{6-18}$$

where S is the number of market segments, which is equal to six in this case. The q_i is the number of VM to be acquired by the customers in the market segment "i." This quantity is decided by the customers' maximum surplus value that is greater than zero for the given price, p, which is offered by a CSP. q_i is a function of a price p.

$$\pi[p] = pQ(p) - C[Q(p)], \quad c_u[Q(p)] \leq p \leq M, \qquad c_u[Q(p)]\, Q(p) = C[Q(p)], \tag{6-19}$$

$$p^* = arg\max_p \pi[p] \tag{6-20}$$

where $Q(p)$ is the summation of $q_i[p]$ of VMs multiplied by the estimated market demand $D_i[p]$ of each market segment. M is the normalized maximum utility value in Table 6—5. This study generalizes this value, which is to make it the same across all the segments ($1.5). $C[Q(p)]$ is the total cost based on the cost assumption of Table 6—4 (internal rationality). In summary, Equation 6-18 is to determine the quantity $q_i[p]$ of VM in each market segment when customer surplus-value is maximum and the total volume $Q(p)$ of VM sales for all market segments. Equation 6-19 is the same as Equation 6-3 with some boundary conditions of price and unit cost. Equation 6-20 is to identify the optimal price for the profit maximization, which is the same as Equation 6-4.

According to customer surplus values, some customers will buy VM others might not buy any for a given price per VM. It is dependent on the type of utility function $U_i\ [j]$ or customers' utility (external rationality) and CSP's offering price p and $c_u[Q(p)]$ per VM (internal rationality), which has been illustrated in Section 6.3.3. The question is, "would it be possible to generate different types of pricing models so that the customers of both the 1st and the 2nd market segments will make a purchasing decision?" For example, if a CSP can offer a specific percentage discount on VM price but customers must purchase VMs in a bulk size? This question leads to creating the bulk-selling pricing model.

6.3.6 Bulk-Selling Pricing Model

In comparison with on-demand, a CSP can generate a bulk-selling or services-bundle pricing model. The goal of the bulk-selling is to encourage cloud customers to buy more for a better pricing deal. There are many examples of the bulk-selling price model, such as one of the retail giants, Costco Wholesale. The telco industry often uses service-bundle for different market segments. Service-bundling means bundling different types of services into one package and bulk-selling is to group different sizes of the same product or service into one package. Two models are closely related.

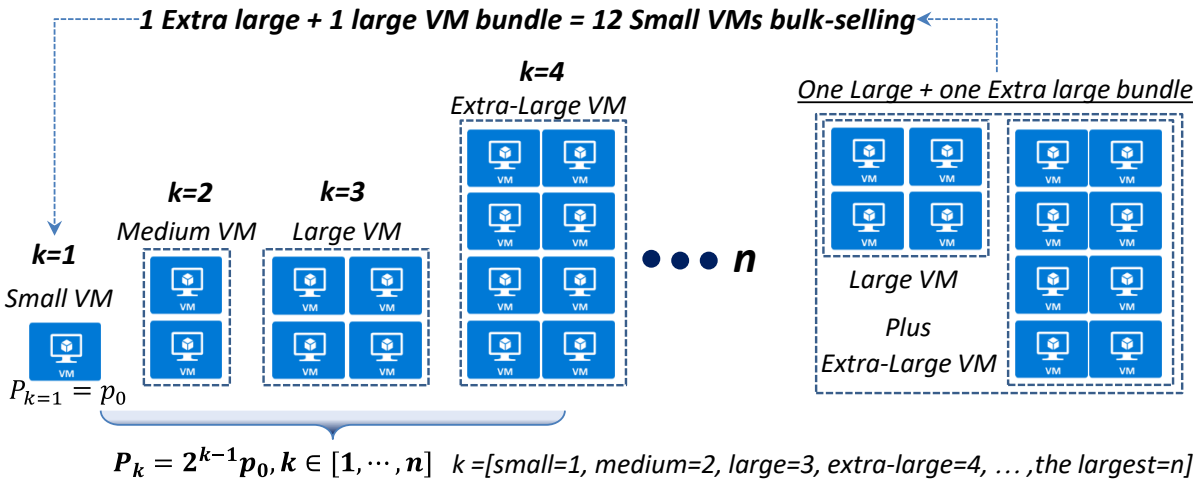


Figure 6—4 Cloud Service Bundle Vs. Bulk-selling Pricing Model

For example, one large and one extra-large size instances can be formed as one package, which is equivalent to 12 small VMs for bulk-selling (see Figure 6—4). According to [117] observations, the AWS price of the current size of the VM is equal to 2 power of "*k*" minus 1 and multiply by the price of the smallest or baseline VM size (where $k = 1,2,\cdots$, the current size of VM) Mathematically, and it can be written as $p_k = 2^{k-1}p_0$ and p_0 is the price of the smallest VM size, p_k is the current size of VM. Such a prices scheme is adopted by many CSPs for their majority types of VMs. The distinct advantage of adopting this pricing scheme is that the CSP can increase capacity flexibility by building a large VM resource pool with finer granularity and minimize a footprint of cloud infrastructure in a cloud data center. The advantage of the service bundle is that it can reduce the investment budget, increase sales, and meet the fluctuation demand for cloud resources.

Both bulk and bundle type of pricing scheme can be tailed for a particular business application, such as mission-critical workload, virtual data center, DR, and collocation, which the CSP will only sell for a fixed number of VMs as a bulk or a clustering package. The cloud customers will decide whether to buy or not, based on their maximum utility (surplus) values. This bulk-selling model can be built equations from 6-21 to 6-26. If the assumption of the bulk-selling size is "*b*," the "mod" function "B" can be adopted to test whether any requested quantity matches the bulk-selling size or not. If it does not match, then a negative value (for example, -200) is set artificially to reflect a customer surplus value, which means to reject the customer's purchasing request (Equation 6-21). Otherwise, the customer's surplus-value will be calculated (Equation 6-22), but it should be greater than zero (Equation 6-23).

$$IF\ B = q - b\left\lfloor\frac{q}{b}\right\rfloor > 0 \rightarrow\ CS_i = -200, \qquad \forall b \in q = \{1,2,\cdots,q_m\} \tag{6-21}$$

$$CS_i[p,q(b)] = \left(\left(\sum_{j=1}^{q} U_i[j]\right) - pq(b)\right), \quad Otherwise, \quad q(b) = nb, \qquad n = 1,2,\cdots \tag{6-22}$$

Then, comparing values in the market segment i and find the maximum surplus-value. Based on this maximum surplus-value, the VM quantity q_i can be identified in the market segment i.

$$q_i[p,b] = q_i : \max(CS_i[p,q(b)]) > 0 \tag{6-23}$$

Multiple market demand D_i with q_i in the market segment i (Equation 6-24) and sum up all quantities of market segments, then we can optimize both price p and b to find the maximum profit value (Equation 6-25 and Equation 6-26)

$$Q(p,b) = \sum_{i=1}^{S} q_i[p,b]D_i[p] \tag{6-24}$$

$$\pi[p,b] = pQ(p,b) - C[Q(p)], \quad c_u[Q(p,b)] \le p \le M, \quad C[Q(p)] = c_u[Q(p,b)]\ Q(p,b) \tag{6-25}$$

$$[p^*, b^*] = \underset{p,\ b}{\operatorname{argmax}}\, \pi[p,b] \tag{6-26}$$

For example, if a CSP offer the bulk size b is 4 and the VM price is \$0.5, the surplus values are set to negative ($CS_i = -200$) for all quantities of q that is not divisible by a package size ($b = 4$). Otherwise, the customer surplus value will be calculated. The maximum surplus-value of the 1st market segment is 1.125, which is corresponding to the VM quantity of 4. There are other VM quantities (8 and 12) that can be divisible by the package size, but the surplus-value is either 0.0076786 or -1.34518. In comparison, purchasing 4 VMs has the maximum surplus values for cloud customers. Base on a similar line of reasoning, the total sales volume for segments can be found, which is equal to $Q(p,b)$ = 7,108 (Equation 6-24). From Equation 6-25 and Equation 6-26, the optimal price point p^* and package size b^* can be found (More details will be covered in the following sections).

The bulk-selling pricing model is just one of the retail pricing strategies. Is it possible to introduce an upfront fee for further VM price reduction? The question leads to a "two-part tariff" pricing model, which is also called the reserved pricing model.

6.3.7 Reserved Pricing Model

The reserved (or two-part tariff) pricing model can be considered a price mixing strategy. It consists of two parts of pricing. This model is widely adopted by many service industries, such as retail, entertainment, airline, and telco. The purpose of this model is to give CSPs more flexibility to target various market segments. The model can be defined as following Equations from 6-27 to 6-32.

$$CS_i[q,p,F] = \left(\sum_{j=1}^{q} U_i[j]\right) - qp - F\,, \quad 0 < F \le F_m \tag{6-27}$$

$$q_i(p,F) = q_i\colon\ max(CS_i[q,p,F]) > 0 \tag{6-28}$$

$$IF\ CS_i[q,p,F] > 0, \qquad cq_i[p,F] = 1, \qquad ELSE\ \ cq_i[p,F] = 0 \tag{6-29}$$

$$Q(p,F) = \sum_{i=1}^{S} q_i(p,F) D_i\,[p], \qquad C[Q(p)] = c_u[Q(p,F)]Q(p,F) \tag{6-30}$$

$$\pi[p,F] = pQ(p,F) - C[Q(p)] + F\sum_{i=1}^{S} cq_i[p,F]D_i[p] \tag{6-31}$$

$$c_u[Q(p,b)] \le p \le M,$$

$$[p^*,F^*] = \underset{p,\ F}{\operatorname{argmax}}\,\pi[p,F] \tag{6-32}$$

where $p^*, F^*$are the optimal price for usage charge and optimal reserved fee respectively and F_m is the maximum fee can be estimated. In this chapter, we can set to \$100. $cq_i[p,F]$, it represents the reserved account for targeted customers in the market segment "i." If the customers' surplus value is less than and equal to zero, it means customers will not pay upfront fee F ($cq_i[p,F] = 0$).

In comparison with bulk-selling, reserved pricing also has two variables. It means that the cloud customers must pay the upfront reserved fee and then they can purchase VM. In return, CSPs offer a significant discount on the usage charge to encourage cloud customers to buy more. The

benefit of this model can boost sales and increase the profit margin. If a CSP would like to increase the profit margin further, the next logical step is to combine both bulk-selling and reserved together.

6.3.8 Bulk-Selling plus Reserved Pricing

This model is to leverage both bulk-selling and two-part tariff models' advantages. However, the benefits of the two models do not have an additive effect. The net profit increase is not bulk selling plus reserved. Very often, the value of the profit margin increases very small because the cloud customer surplus-value may approach its bounded limit in the separated models. The following Equations from 6-33 to 6-39 represent this model.

$$\begin{gathered} IF\ B = q - b\left\lfloor\frac{q}{b}\right\rfloor > 0\ \rightarrow\ CS_i = -200, \\ \forall b \in q = \{1,2,\cdots,q_m\}, \qquad q(b) = nb, \qquad n = 1,2,\cdots \end{gathered} \tag{6-33}$$

$$\begin{gathered} Otherwise, \\ CS_i[p,q(b),F] = \left(\left(\sum_{j=1}^{q} U_i[j]\right) - pq(b) - F\right), \qquad 0 < F \le F_m \end{gathered} \tag{6-34}$$

$$q_i[p,b,F] = q_i : \max(CS_i[p,q(b),F]) > 0 \tag{6-35}$$

$$Q(\cdot) = Q(p,b,F) = \sum_{i=1}^{S} q_i(p,b,F) D_i\,[p] \tag{6-36}$$

$$IF\ CS_i[p,q(b),F] > 0, \quad cq_i[p,b,F] = 1, \quad ELSE \quad cq_i[p,b,F] = 0 \tag{6-37}$$

$$\begin{gathered} \pi[p,b,F] = pQ(p) - C[Q(p)] + F\sum_{i=1}^{S} cq_i[p,b,F] D_i\,[p] \\ c_u\big(Q(p)\big) \le p \le M\,, \qquad C[Q(p)] = c_u\big(Q(p)\big)Q(p) \end{gathered} \tag{6-38}$$

$$[p^*, b^*, F^*] = \underset{p,b,F}{\operatorname{argmax}}\, \pi[p,b,F] \tag{6-39}$$

The goal of this model is to maximize the profit margin by bulk-selling to encourage the customers to buy more VMs and by the upfront reserved fee to motivate the cloud customers to consume more for less unit cost per VM. In comparison with other models, this model has three

variables to be optimized. Now, the question is how to optimize these pricing variables for profit maximization, in which the question has been left unanswered in Section 6.3.3.

6.4 Genetic Algorithm and GA Parameters Setting

6.4.1 Proposed Methods

There are many possible optimization methodologies or techniques that can be applied for the optimizing problem, such as gradient descent, Genetic Algorithm (GA), and simulated annealing. Gradient descent cannot be applied because the profit equation is non-contiguous. Simulated Annealing could be one of the potential methods for this solution because it usually is better than greedy algorithms, but the technique can be slow, especially if the cost function is expensive to compute. Subsequently, this study adopts GA to solve this problem. It can be solved very quickly (30 seconds/per each iteration if without further improvement).

6.4.2 Genetic Algorithm

The useful properties of GA are 1.) It does not require specifying an objective function, 2.) The objective function does not have to be either continuous or linear, 3.) It takes less computational memory, 4.) It can optimize multiple variables in parallel, and 5.) Some local optimal solution could bring some insights as to potential price options to form a pricing strategy. The basic idea of evolution computation strategy is "trial and error" is shown in Figure 6—5 and Figure 6—6. The principle of this method is based on the underlying microevolution of both mutation and natural selection [266], which is to mimic the biological process that is searching for an optimal solution in a problem domain.

Based on Equation 6-3 and 6-4, the goal of this research is to find the maximum value of the profit "π" by searching for the optimal price point "p." We know that price, cost, and sales quantity are interdependent. It is challenging to define some precise sub-functions for the solution. However, we can set up price p as "genes" and let a set of price, quantity, and unit cost to be a chromosome (a set of parameters for the solution). A string of chromosomes is called the genome. The entire combination of prices (genes) is called genotype, and the corresponding profits are called phenotype, as shown in Figure 6—5. Note that the optimal variables can be extended to bulk-selling size "b" or upfront fee "F." In Figure. 6—5 and Figure 6—6, we only show the optimal price and bulk size.

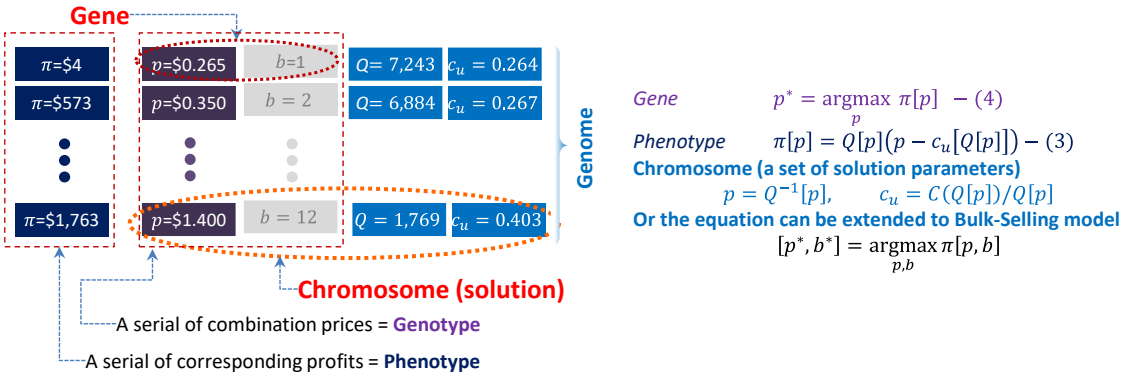


Figure 6—5 Details of GA Calculation for Maximum Profit π for On-demand

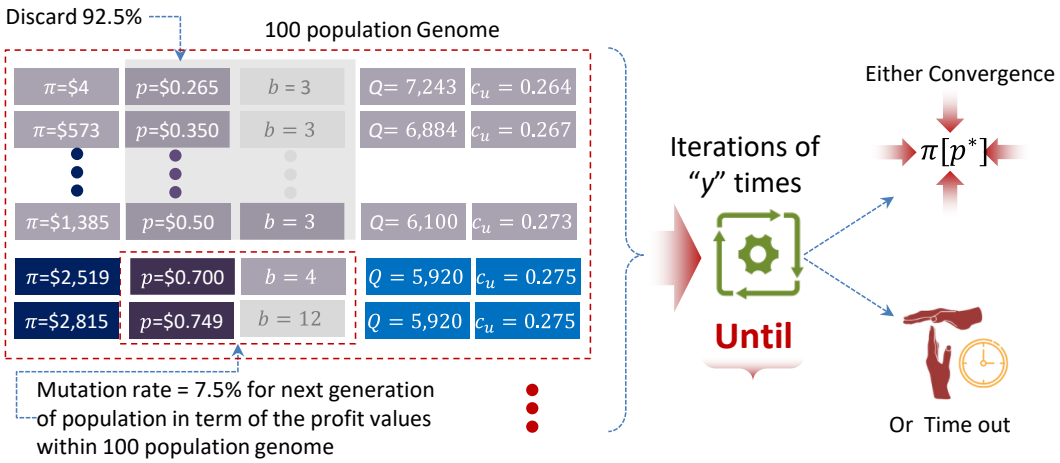


Figure 6—6 Performance Function and Criteria of the GA Solution

In the following example of the on-demand pricing model, this chapter assumes the price value in the range [0, $1.5] because no customers will expect to buy the VM more than the maximum amount of their utility value. If we first trial the initial price value or a gene as p = $0.265, we should have the profit $\pi[p]$ = $4, unit cost $c_u[Q[0.265]]$ = 0.264, and the total sales quantity Q =7,243. Clearly, it is not an optimal price. So, we let GA compute Equation 6-4.

For each of 100 population size (A "standard GA" parameter of the population size can be set up between 100 and 200 [312]), the best 7.5% of prices p or genes will be kept and discard 92.5% in term of better profit values because we set up the mutation rate is 7.5% in this experiment process. After "y" times of this iteration, we can find the maximum value of profit based on the performance of the convergence resolution or stopping condition for GA is either r_{con}= 0.01%

(Equation 6-40) or time out = 30 seconds (roughly between 280-350 GA iterations) (See Figure 6—6).

$$r_{con} = \left|\frac{\pi[p_{m+1}] - \pi[p_m]}{\pi[p_m]}\right| < 0.01\%, \qquad m = 1, \cdots, N \tag{6-40}$$

where $\pi[p_m]$ is profit estimated at iteration m with price p_m.

6.3.3 Experiment Implementation and A Pseudocode

. A Pseudocode is presented to articulate this genetic algorithm process as Algorithm 1. To carry out this iterative process, we can adopt different software applications to implement our experiments, such as Matlab, R and even Microsoft Excel Solver. R has two convenient packages: GA and Genalg, which can quickly run our tests. The input data for the tests are sourced from Table 6—5 as initialized parameters. The outputs are the optimal values of four pricing models for on-demand, bulk-selling, reserved, and bulk + reserved. To simplify the experiment, the variation of the market demands is excluded in terms of price variation in each market segment.

Algorithm 1: Pseudocode of Cloud Pricing Models

PROGRAM: Genetic Algorithm for CloudPricingModel

Input: PopulationSize $N(m \leftarrow 1 \cdots 100)$, **PriceRange** $R \in [0, K_i]$, **CrossoverRate** $C_r \leftarrow 0.6$, **MutationRate** $m_r \leftarrow 0.075$ // Initalize Parameters;

Output: $p^* \leftarrow arg\max_p \pi[p]$ // Find the Optimal Price p for Cloud Business Profit Maximization;

$P_0\{p_m\} \leftarrow \{p_1, p_2 \cdots p_m\} \in [0,\ K_i]$ **InitalizePopulation** // Randomly Select p_m : Population size, Problem Size;

$\pi[p]\ \leftarrow Q[p] * (p - c_u)$ **Objective Function** // Calculate Objective Function;

$\frac{\pi[p_{m+1}]-\pi[p_m]}{\pi[p_m]}$ **EvaluationPopulation** // Use Fitness Function for Evolution Initial Population $P_0\{p_m\}$;

$\pi[p] \leftarrow\ p$ **GetBestSolution** // Assign the Best Price to the Object Function from Initial Population $P_0\{p_m\}$;

While != **StopCondition** (Time ≤ 30 sec without improvement) OR ($r_{con} < 0.01\%$) DO // either Time Less Than 30 secs or $\pi[p]$ Convergence

$P_g\{p_m\} \leftarrow P_0\{p_m\}$ **SelectParents** // $P_g\{p_m\}$ // Select Parents Population;

$cg \leftarrow 0$ **SetToZero** // Sign Children Generation to Zero;

FOREACH $P_g1, P_g2 \in\ P_g$ **DO** // Iteration Process

$P_{cg}1, P_{cg}2\ \leftarrow$ **Crossover**$(P_g1, P_g2,\ C_r)$ //Perform Crossover and Sign to Children Population;

$P_{cg}\ \leftarrow$ **Mutation** $(P_{cg}1, m_r)$ //Perform Mutation;

$P_{cg} \leftarrow$ **Mutation** $(P_{cg}2, m_r)$ //Perform Mutation;

ENDFOR

EvaluatePopulation (P_{cg}) // Use Fitness Function to Evaluate Children Population P_{cg};

$\pi[p] \leftarrow\ p$ **GetBestSolution** (P_{cg}) // Assign the Best Price to the Object Function from Children Population P_{cg};

$P_{cg} \leftarrow P_{cg}$ **Replace** (Population, P_{cg}) //Insert Offspring;

$cg \leftarrow cg + 1$ // Create New Generation;

ENDWHILE

Return $p^* \leftarrow arg\max_p \pi[p]$;

6.5 Experiment Results

6.5.1 On-Demand Pricing Model Results

Table 6—6 shows the final results for all pricing models, including on-demand, which CSPs should charge $0.749 per VM instance/hour for the maximum profit of $2,815. The average unit cost is about $0.269. In comparison with cost-based pricing, the on-demand price can boost a 73% profit margin if we take account of the external rationality. Although the profit margin (100%) of the cost-based pricing looks very attractive, it is not optimal. The result of this comparison means the cost-based price is significantly underestimated the unit price of cloud customers' willingness to pay.

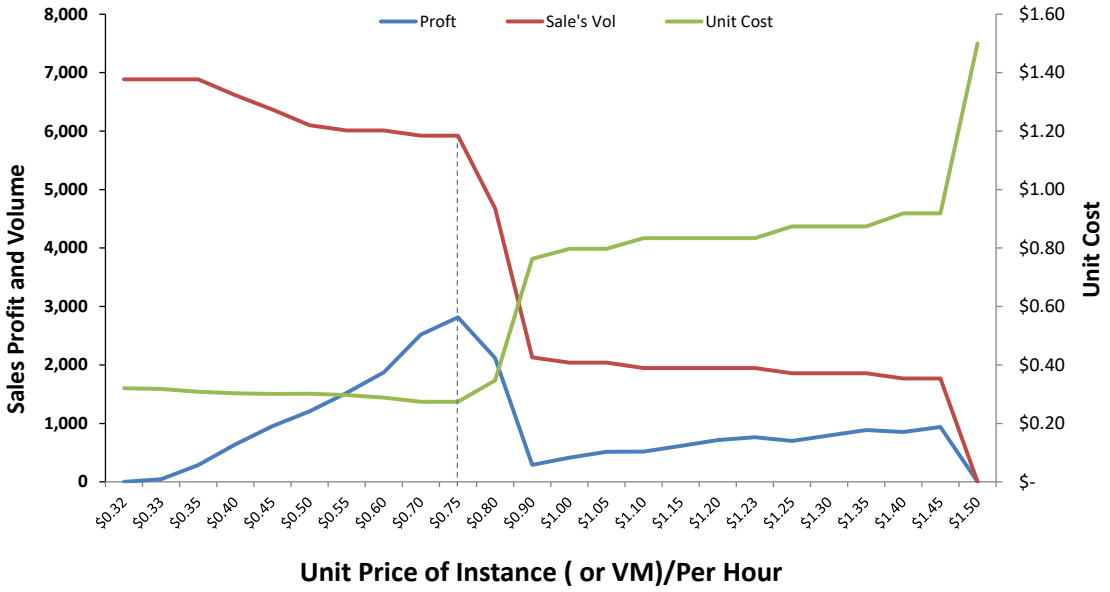


Figure 6—7 On-Demand Price Model of Price Change

In Figure 6—7, this chapter shows how the profit is evolving in terms of offering prices. There are a few local-optimal prices, such as $1.225 $1.450. To overcome these local-optimal values, different initial values of prices can be tested. As section 6.3.4 indicated, the on-demand pricing model is just one of the price models for various market segments. Other models, such as bulk-sell, are possible for CSPs to gain a higher profit margin.

Table 6—6 The Result of On-Demand Pricing Model

Pricing Models	Type of Pricing Models	Optimal price: p	Optimal Bulk Size	Reserved Fee F	Unit Cost: c_u	Total Cost C	Total Sales Quantity: q	Total Revenue: R	Maximum Profit: π	Profit Margin
Markup	Cost-based	$0.533	NA	NA	$0.266	$1,625	6,100	$3,250	$1,625	100%
On-demand	Value-based	$0.749	NA	NA	$0.274	$1,625	5,920	$4,440	$2,815	173%
Bulk-Selling	Value-Based	$0.675	4	NA	$0.265	$1,886	7,108	$4,798	$2,912	179%
Reserved Fee	Value-based	$0.279	0	$5.572	$0.277	$1,566	5,652	$5,019	$3,465	213%
Bulk+Reserved	Value-based	$0.587	12	$1.958	$0.264	$1,935	7,332	$5,499	$3,564	219%

Note: Bulk-Selling Price is not based on the maximum profit but a 10% discount on the on-demand price.

6.5.2 Bulk-Selling Model Results

The bulk-selling price model requires optimizing two variables. One is the selling price, and the other is the bulk-selling size. Based on Equations from 6—18 to 6—23, the cloud customers will only make a purchase decision when their surplus values are higher than their cost. The experiment results show that the package size is 12 and the optimal price is just slightly below the on-demand price or $0.7452 for CSP to achieve the maximum profit margin of 217%. This price will not attract customers to buy in bulk. Subsequently, CSP can give a 10% discount off the on-demand price, which is to set the selling price at $0.675 and reduce the package size from 12 to 4. Even so, the CSP can still achieve a 179% profit margin. If we keep the package size is 4 and give a 10% discount off the on-demand VM, we can plot out the profit evolution along with the price changing, as shown in Figure 6—8.

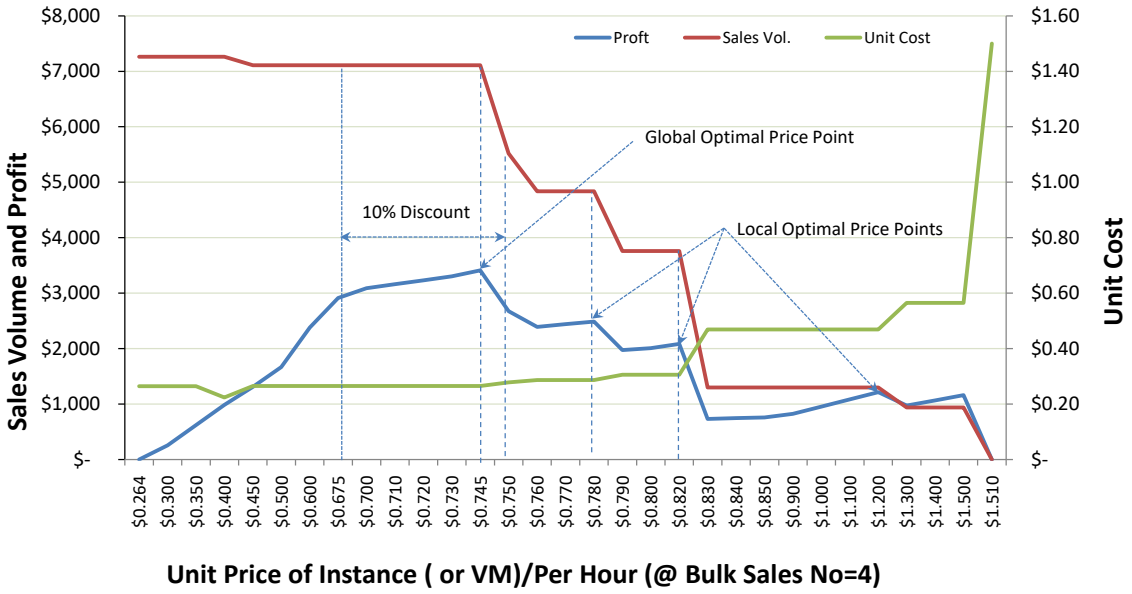


Figure 6—8 Bulk-Selling of Price Change For All Optimized Parameters (BulkSize@4)

If we keep the 10% discount price unchanged and make the variation of the bulk-size from 1 to 12, we can find bulk-size-4 is the local optimal and bulk-size 12 is the global optimal value. These optimal price points (as shown in Figure 6—9) provide more options for CSPs' pricing strategy

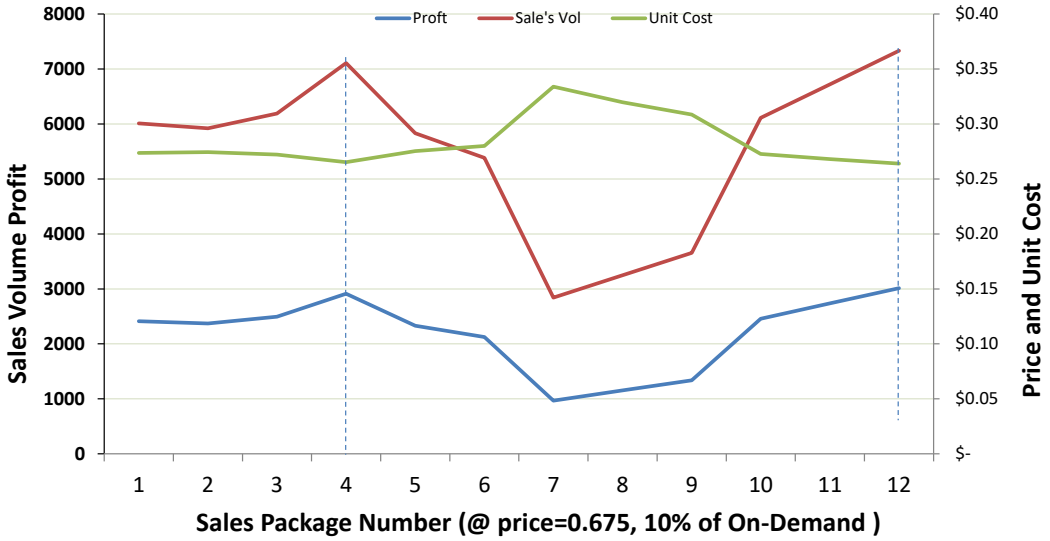


Figure 6—9 Bulk-Selling Package Size Evolution

Intuitively, the downside of the bulk-selling is that some cloud customers do not want to scarify their flexibility of Pay as You Go (PAYG) and still demand a competitive price because their business might not require the bulk-size of VMs. As a result, customers might switch to other cloud competitors. Adopting one price model could cause a CSP to lose its market share if the CSP insists on the bulk-selling model. If a CSP would like to keep both higher profit margin and market share, what is an alternative?

6.5.3 Reserved Price Model Results

The possible solution is a reserved pricing model. The experiment result shows that the reserved price model can achieve a profit margin of 213%. The primary profit contribution is due to the reserved fee, which is $5.572 per account or $3,456. The VM price is $0.279, which is very close to the unit cost, which $0.277. If CSPs keep the reserved fees the same ($5.572) and changing the VM price, we can see the cloud price evolution (See in Figure 6—10).

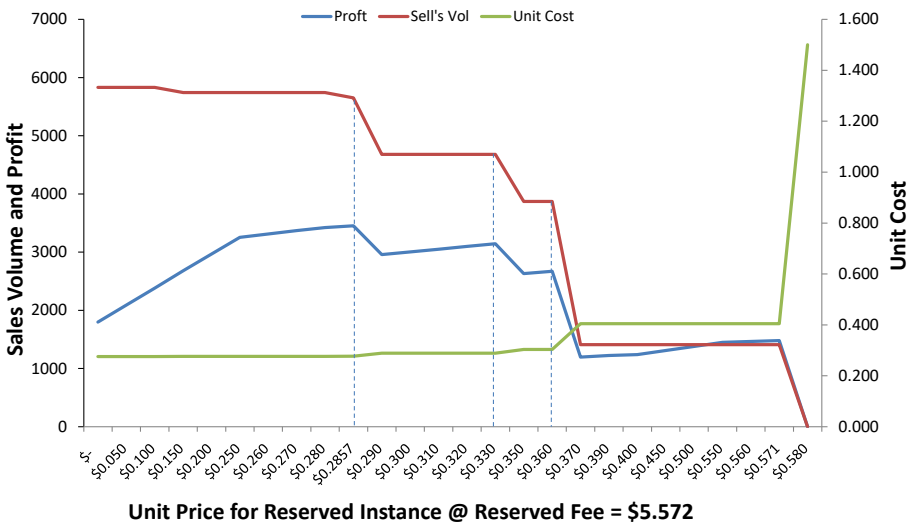


Figure 6—10 Reserved Model of Price Change for All Optimized Parameters @ F=$5.57

Again, if the VM price is kept the same (at $0.279 per VM) and the reserved fee is changed, there will be two local-optimal prices at $1.96 and $6.00 shown in Figure 6—11

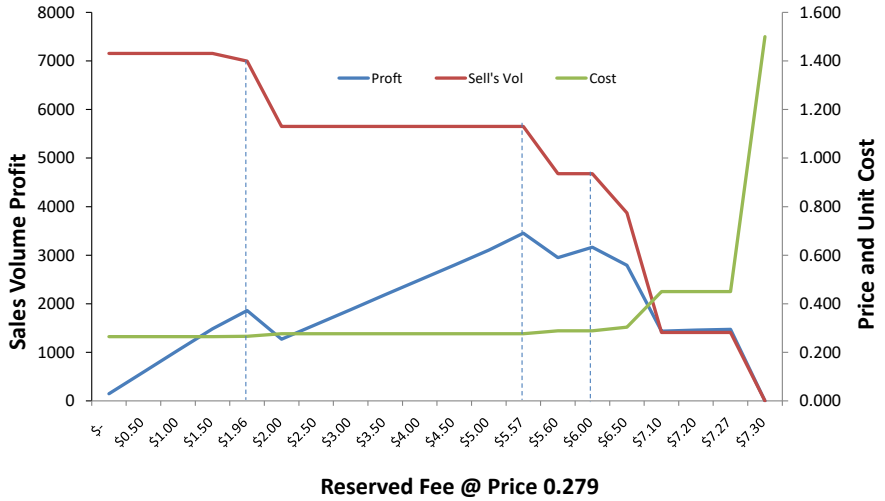


Figure 6—11 Reserved Fee Changing at Price@$0.279

6.5.4 Results of Bulk plus Reserved Pricing

If CSPs would like to increase profit margin further, they can combine both bulk-selling and reserved models. In comparison with a pure reserved model, "bulk + reserved" can grow about 7% profit margin. This model offers different alternatives for CSPs to form a pricing strategy to meet various requirements in different market segments, in which a CSP can increase the usage

charge and decrease the reserved fee or vise verse. The plot of profit, sales' volume and unit cost along with VM price change can be considered as a combined effect of bulk-selling plus reserved as observed in Figure 6—12

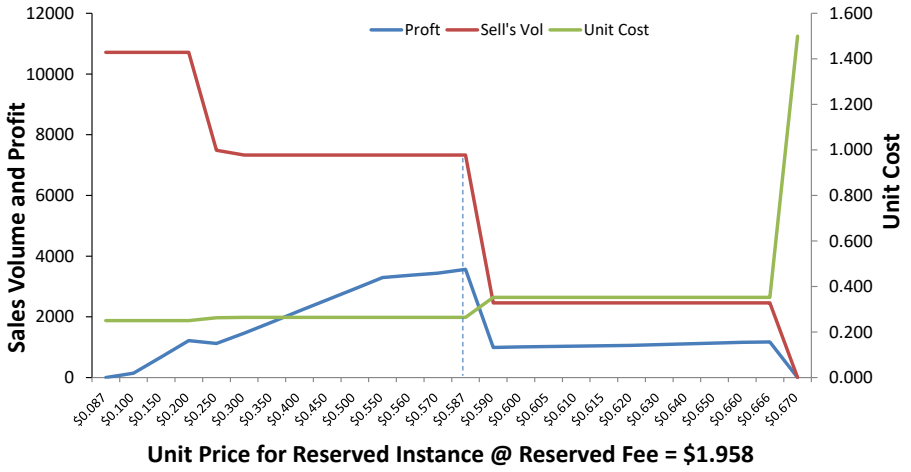


Figure 6—12 Bulk-selling Plus Reserved of Price Change (Fee@$1.958 Bulk Size@12)

Following a similar principle, we can also plot the fee change while the unit price ($0.587) and bulk size (12) are kept the same. The result is shown in Figure 6—13. As we should see, the shapes of the two plots are very similar except the sales volume. There are a few local optimal price points. Again, these price points provide different price options for CSPs to articulate different price strategies.

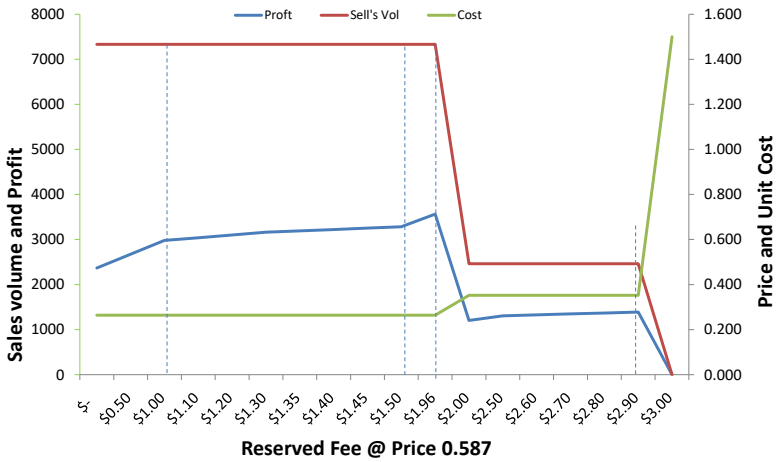


Figure 6—13 Fee Changing of Bulk + Reserved (Price @$0.587, Bulk size @12)

Overall, the experimental results show that the on-demand pricing model can significantly increase CSPs' profit margin if the cloud customers' utility values are included in comparison with the cost-based pricing. The bulk-selling price model is aiming to encourage customers to buy more for fewer usage charges. The reserved pricing model is to decrease more usage charges with the upfront reserved fee. This flexible option can help CSP to maintain a healthy profit margin while the usage price is very competitive. The bulk + reserved model is to provide different options for cloud pricing strategies to maximize the CSP's profit while they can target cloud market segments.

6.6 Analysis and Discussion

This chapter demonstrates a comprehensive framework of how to formulate four value-based cloud pricing models from a value co-creation perspective. In contrast to previous researches that assumed a uniform market with only one utility function, this solution of cloud pricing is much realistic and practical because market segmentation practice has been widely applied to many service industries. The cloud industry is not exceptional. AWS has up to seven different types of pricing models (spot, on-demand, reserved, bare-metal, dedicated host, and Code on Demand) for different market segments. Based on multiple market segments, the GA can find the optimal pricing solution for each model.

6.6.1. GA Performance Evaluation

In comparison with other optimal solutions, the GA process requires less computing memory and power and does not need to specify sub-functions. The object function does not have to be differentiable. It can be either a continuous function or a discrete one. Many software packages can implement the GA process. Even MS Excel Solver can implement it, which is very handy for many practitioners to generate pricing options and form a better and competitive pricing strategy. The GA process can also be updated quickly if the cloud market environment has been changed.

To evaluate the performance of the GA process for the optimal pricing value, we tune one of the GA's parameters: mutation rate into different values and to see which we can achieve a better performance result quickly. According to [323] [324], we applied the mutation rates between 0.001 and 0.5. Our result shows when the mutation rate is equal to 0.075, the profit margin of reserved pricing models can be quickly converged to the maximum value (as shown in Figure

6—14) within the specified timeframe of 30 seconds with 100 population size and converged rate of 0.01%.

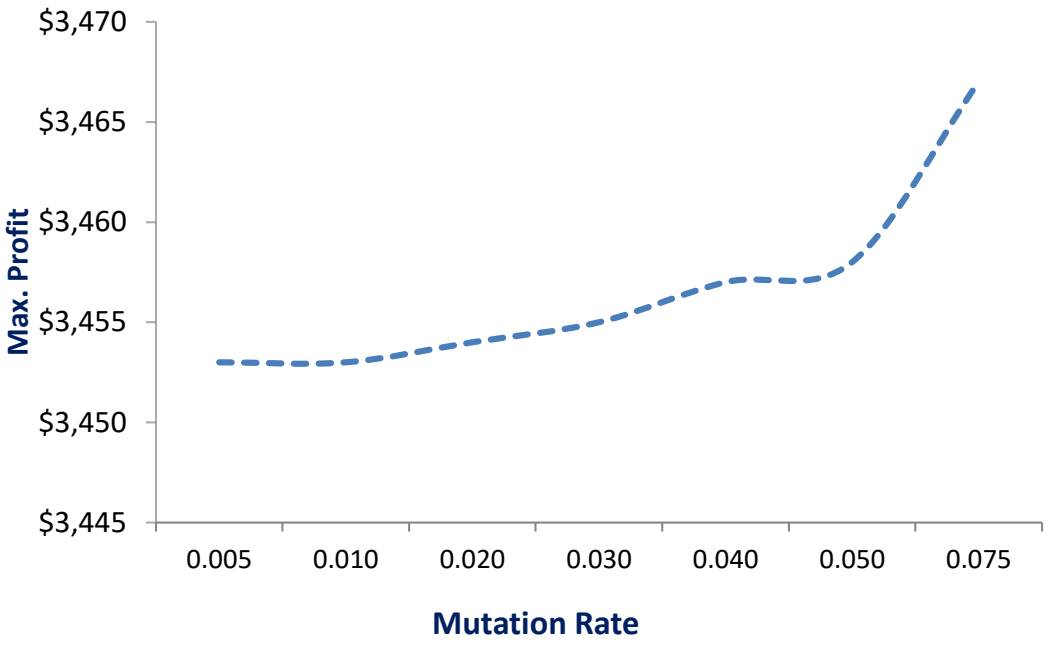


Figure 6—14 GA Performance Evaluation for Different Mutation Rate

6.6.2. Comparison with Created Pricing Models

From the above experiment results, this chapter illustrated that cost-based pricing has the lowest profit margin ($1,625). As Nagle et al. [10] indicated, although the model carries a financial legitimacy, it only provides "mediocre financial performance." Although a 100% profit margin seems to be very attractive, it is "mediocre." The critical issue of cost-based pricing is that it does not include external rationality

On the other hand, on-demand can achieve a higher profit margin and higher sales volume in comparing with cost-based pricing. However, the on-demand model might work well with one business application (or one market segment), but not fit with others. To solve this issue, this study adds more pricing options in a decision-makers' toolbox, both bulk-selling and reserved pricing model can also play their roles in the segmented market. One of the advantages of bulk-selling and reserved models is that they can provide business certainty for cloud resource capacity. The downside is that cloud customers could lose some flexibility. If we compare all four cloud pricing models, the reserved + bulk-selling pricing model can achieve the highest profit margin for CSPs, as shown in Figure. 6—15

In order to gain a higher profit margin, the bulk-selling price model is one of the good pricing strategies, which have been observed in the cloud pricing practice. In fact, bulk-selling is equivalent to AWS, Azure, IBM Cloud, and Google Cloud Platform's reserved instance (without an upfront fee). The only difference is time. With bulk-selling, cloud customers must consume all purchased resources at once. In contrast, AWS, Azure, IBM cloud, or GCP's reserved instances can be consumed from one or three years. The longer the time of cloud resource reservations, the cheaper the unit price is. The reservation time is equivalent to a bulk-size. If we put time and cloud assets depreciation factors aside, the currently reserved instances offered by major CSPs are similar to the bulk-selling or bundle pricing model. As this chapter shown in the experiments, the bulk-selling price model can improve CSP's profit margin by 6% even with a 10% price discount in comparison with the on-demand price. (See Table 6—6 and Figure 6—15). That is why many CSPs encourage cloud customers to adopt reserved (or bulk-selling) instances with a discount price.

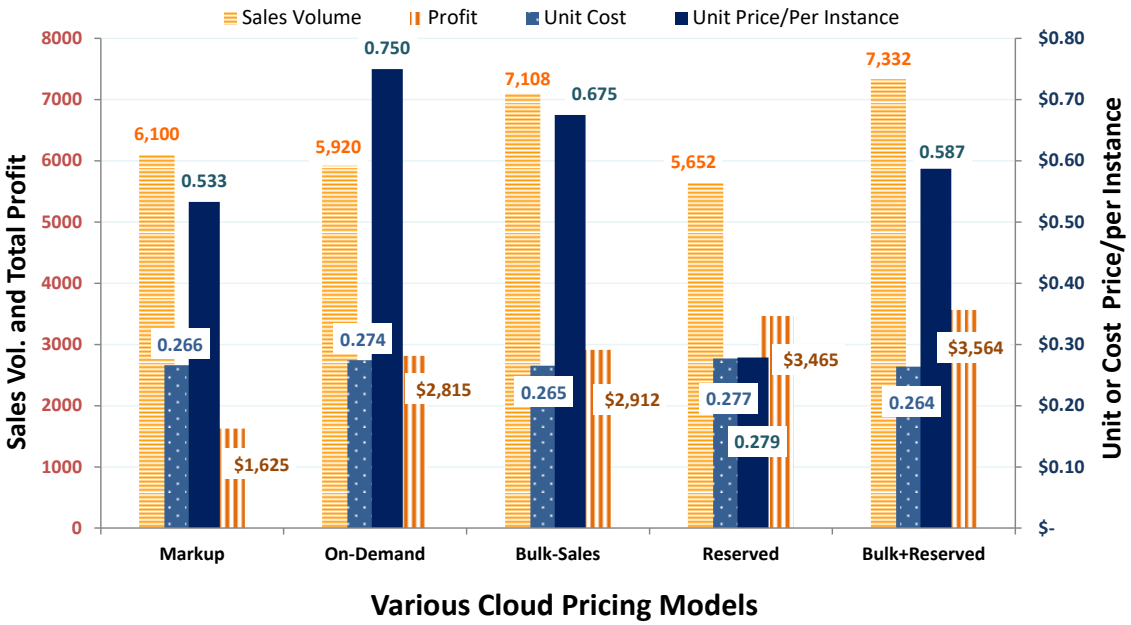


Figure 6—15 Comparison of Different Pricing Models with Six Market Segments

If CSPs would like to make further improvements in their profit margin, they can introduce the upfront reserved fee (two-part tariff) and reduce the usage charge of VM unit price in return. With the upfront reserved fee, the CSPs can reduce VM usage charges as low as the production cost and still maintain a healthy profit margin, which is around 213%. Comparing with the "on-demand" model, the usage charge (or unit price of VM) drops nearly 63%. Now, the upfront

reserved fee becomes the major profit contributor to CSP's profit. If the cloud customers are not willing to pay a higher upfront reserved fee, CSPs can adopt the mixing model of bulk + reserved fee. The above experiment result shows that by a combination of the bulk and upfront reserved fee, CSPs can lift profit margin by 219% and reduce the upfront fee by 65% (in comparison with pure reserved model) and decrease VM price by 22% (in comparison with on-demand).

Checking the sales volume of VM across all segments, the customers in segment 6 would not purchase any number of VM for proposed price models, as shown in Table 6—7. This is because their utility function is risk-taking. The shape of the utility function is concave. None of the proposed pricing models would capture customers' surplus values in segment 6 unless a CSP can offer a substantial discount, such as a 60% - 80% price reduction of the on-demand.

Having a considerable price discount for one market segment alone and scarifying other markets' values is not a good pricing strategy because the cloud business profit will decline significantly. For example, if the price is dropped by 60% across all market segments, the profit margin will be reduced by about 84%. Selling cloud resources with a substantial discount price is not a sustainable business practice for CSPs.

Table 6—7 Sales Volume of VM for Each Model/per Cloud Customer

Pricing Models	Seg 1	Seg 2	Seg 3	Seg 4	Seg 5	Seg 6	Total
Cost-based	2	12	8	6	12	0	40
On-Demand	2	12	6	6	12	0	38
Bulk-Selling	4	12	8	8	12	0	44
Reserved Fee	0	12	9	6	12	0	39
Bulk + Reserved	0	12	12	12	12	0	48
Combine Revenue	$402.96	$1,660.50	$486.00	$1715.50	$656.10	0	$4921.06
Combine Cost	$147.41	$651.90	$190.80	$390.57	$257.58	0	$1638.26
Combine Profit	$255.55	$1,008.60	$295.20	$1,324.93	$398.52	0	**$3,282.80**
Preferred Price models	On-demand	Bulk	Bulk	Reserved Fee	Bulk	Discount Pricing Model	

However, what this study has observed is that many leading CSPs still offer a discount price, such as spot instance, preemptible and low priority, for risk-taking customers. The reason that CSPs can offer a massive discount without cannibalizing the profits from other market segments is that the cloud service with a discount price has many restricted conditions, such as preemptible, time limit, limited availability zone, etc.

From a marketing perspective, the spot or preemptible instance is more like the "Razor-and-Blades" pricing strategy [313], which is to use a lower price to simulate customer's demand. Practically, it would not be a good idea for business customers to rely on spot or preemptible instances (VMs) alone for a mission-critical application, although the price of spot instance is very competitive.

Table 6—7 shows that the customers in segment 4 will purchase more than what they need if a CSP offers bulk-selling price models only. However, if all price models are released spontaneously in a cloud market, customers of segment 1 will acquire VM according to the on-demand price model because it has the highest surplus value and the lowest cost, which we assume the lowest unit price as a measurement. Customers of segment 2 will adopt the bulk-selling model. Customers of segment 3 will also select the bulk-selling model. Customers of segment 4 will choose the reserved fee price model. Customers of segment 5 will be the same as the segment 2. Customers of segment 6 will not buy any VM resources. The sum of six market segments for all value-based price models is also 40, which is the same quantity as the cost-based pricing model. However, the average profit margin is over 200%, and the total cost is just slightly increasing by 0.8% in comparison with cost-based pricing.

These value-based price models provide a wide range of pricing options for CSP to achieve the maximum profit by capturing more customers' surplus values from various market segments. Based on the market segmentation theory [170], the ideal strategy for CSPs is to have personalized pricing because the better the information about the customers, the fine partition of the customers into a group and the larger the possibilities for CSP to extract customer surplus." In other words, the one-price model is dedicated to one customer, which is also known as the 1st order price discrimination. However, it would be impossible for CSP to implement a personalized pricing strategy because it requires a lot of managerial and sales resources. The alternative solution is "market segmentation." Naturally, different market segments will lead to different utility values. It results in various price models with multiple optimal price points to meet different preferences. Table 6—8 provides summary information of all models that we have proposed in this work in terms of different application scenarios, advantages, and disadvantages.

Table 6—8 Summary of All Pricing Models

Pricing Strategy	Pricing Models	Application Scenarios	Advantages	Disadvantages
Cost-Based	Cost-Based Pricing	Enterprise internal cost modeling	Recover the cost bottom line	Arbitrary
Value-Based	On-Demand	Application develop, solution architecture	Flexible	High Cost
	Bulk-Selling	Long-term Web hosting required large server clusters, Deliver SLA App.	Having a certain % price discount	Have to buy in a bulk size
	Reserved	Having cloud resource certainty	Relative lower cost	Lack of flexibility
	Bulk + Reserved	Large server clustering & resource certainty	Min. cost	Lowest flexibility

6.6.3. Comparison with Other Works

To the best of knowledge, there has been no research work that has been done to assume multiple market segments for Cloudonomics. Although some previous works [80] claimed that the uniform price would not suffer any revenue loss in comparison with the 1st order price discrimination, this chapter illustrated this claim was contradicting the theory of market segmentation [170] [172]. Based on the simulation result, this chapter has demonstrated that if there is only one market segment defined by one utility function that is assumed to be an iso-elasticity or linear or exponential utility, the profit loss will be from 12.15% up to 84.51% and the revenue loss will be from 14.30% up to 79.93% as shown in Table 6—9.

Table 6—9 Profit, Revenue and Optimal Price Comparison with Other Works

Sources	Six Segments	Uniform Market[113][118][80]	Uniform Market [55]	Uniform Market [275]	Uniform Market [69]	Uniform Market [252]
Utility	Six Utility Functions	Iso-Elastic Utility, $\alpha<1$	Iso-Elastic Utility, $\alpha=1$	Iso-Elastic Utility, $\alpha>1$	Linear by Diminish Return	Exponential Utility
Equivlent Utility Function	$U_i(q), i = 1 \cdots 6$	$U(q) = K\frac{q^{1-\alpha}-1}{1-\alpha}$	$U(q) = K$	$U(q) = K\frac{q^{1-\alpha}-1}{1-\alpha}$	$U(q) = U_0 - \alpha p$	$U(q) = K(1-e^{-\alpha q})$
Optimal Cost	\$0.27	\$0.29	\$0.35	\$0.27	\$0.37	\$0.77
Optimal Price	\$0.75	\$0.83	\$0.81	\$0.41	\$0.96	\$1.50
Max.Profit	\$2,815	\$2,473	\$2,449	\$939	\$1,240	\$436
Max. Revenue	\$4,440	\$3,805	\$3,802	\$2,697	\$2,033	\$891
Profit loss	0%	12.15%	13.00%	66.64%	55.95%	84.51%
Revenue loss	0%	14.30%	14.37%	39.26%	54.21%	79.93%
Sales Vol.	5,920	4,589	4,679	6,525	2,039	594

In summary, the value-based price modeling, together with the comprehensive pricing framework, is better than the current state of the art of cloud price modeling, which has been highlighted in Table 6—3.

6.7 Summary

This chapter has demonstrated the way how to formalize four value-based cloud pricing models based on both internal rationalities of CSP's costs and external rationality of customers' utility values and market segmentation. In comparison with pure internal rationality of resource or cost-based pricing models, this approach to cloud price modeling is practical and straightforward if customers' utility functions and market segments have been defined. Moreover, optimizing GA requires less computing memory and power and does not need to specify sub-functions in comparison with other methods. With GA, these pricing models can also be updated quickly. The significance of this work is to establish a comprehensive framework for practical solutions to estimate and estimate optimal cloud prices.

Chapter 7

Conclusions, Discussion and Future Directions

This final chapter provides the conclusions of this thesis, which are underpinned by the key findings and main contributions. Moreover, it discusses many open challenges and future directions in cloud price modeling due to the rapid development of the cloud business and cloud engineering ecosystem, which is beyond the reach of this thesis.

7.1 Conclusions and Discussion

Cloud pricing is moving away from a physical box-oriented model to a virtual machine-based model, and then to an abstract sandbox-based model. Many CSPs are starting to offer cloud pricing based on an abstract level of software. To some extent, the pricing of the serverless sandbox can be considered as modeling No Operation Systems (No OS or NoOps), which is an evolutionary direction from a pure development environment to an integrated environment of both development and operation or DevOps.

However, this does not mean that cloud users can ignore the underlying cloud infrastructure, such as cloud security, workload balancing, horizontal or vertical scaling, auto-failover or high availability, and disaster recovery. All these cloud features will be part of a CSP's responsibility. They become a part of SLA measurement or service-based pricing. Cloud users do not have to

---

This chapter is mainly derived from:

- **Caesar Wu**, Rajkumar Buyya, and Kotagiri Ramamohanarao, "Cloud Pricing Models: Taxonomy, Survey and Interdisciplinary Challenges," ACM Computing Surveys, Volume 52, No. 6, Article No. 108, Pages: 1-36, ISSN 0360-0300, ACM Press, New York, USA, October 2019

get their hands dirty to tune these cloud features directly. They only need to automate and monitor them and make sure they can be delivered. This is why Kubernetes, Apache Mesos and Docker Swarm have emerged as essential tools behind the transformation of cloud pricing.

As a result of this evolution, we can see that each CSP often leverages its business application strengths to optimize its cloud pricing model. Based on our simple observation, we conclude that AWS can be seen as providing online retail-oriented pricing for its cloud services. Azure is software-oriented pricing, and GCP is search engine optimization (SEO) oriented pricing. Other CSP competitors can leverage different application strength for their cloud pricing, such as e-commerce, utility services, healthcare, cyber-security, and Supervisory Control and Data Acquisition (SCADA). The details of the cloud price modeling problem and objectives were discussed in Chapter 1.

Chapter 2 gave a full description of taxonomy and a comprehensive survey of cloud price modeling in terms of cost-based, market-based, and value-based pricing strategies. It also provided the detail categories for 60 pricing models. Based on that classification, Chapter 2 reviewed 10 of the current state-of-the-art models regarding their pros and cons.

Chapter 3 discussed one of the two themes of this thesis. It addressed the issue of how to price new cloud service features or characteristics. It covered two topics. One is how to price the ever-growing new features of cloud service. The other is how to calculate the depreciation rate of cloud computing due to Moore's law. The novel solution is to introduce the concept of G.E. Moore's intrinsic and extrinsic values of cloud services. With the hedonic analysis, these two issues can be quickly resolved, with the extrinsic value of cloud service being about 43% above the baseline service price. In addition, if we adopt AWS's panel data, this chapter showed that the cloud service pricing deprecation rate is 20%, which is at a far slower pace than computer hardware. This is due to the new and innovative cloud service features.

Chapter 4 introduced another theme of cloud service pricing, which is to focus on the baseline service pricing. It begins with the theory of market segmentation. Based on the classical segmentation theory, this chapter discussed a hierarchical clustering method to extract the cloud business customer usage patterns from Google's public dataset. With these usage parameters and AMD's virtual machine guidelines, the chapter aligned the usage patterns with a particular business application. Then it used the Time Series method to predict market demand. The final step of the cloud market segmentation is to combine the extract usage patterns with predicted

cloud market demand. Ultimately, Chapter 4 built one of the foundations for value-based cloud price modeling.

Chapter 5 focused on how to establish various cloud customers' utility functions for their business applications in terms of their resource consumption, in which a number of virtual machines are required. The chapter articulated six types of cloud utility functions for six types of market segments that are aligned with various cloud business applications, being web hosting or content delivery, virtual desktop infrastructure (VDI), mission-critical workload, e-commerce online checkout system, disaster recovery (DR) and backend data process workloads.

In particular, the utility functions of cloud customers are built upon Markov analysis, queueing theory, alpha-fair utility network utility, risk-averse, risk-taker, and an additive relationship. This chapter created another cornerstone for the next chapter of value-based cloud price modeling.

Chapter 6 discussed the final step of value-based cloud price modeling for baseline cloud services. It provided four types of cloud price models that are very common across many retail and service industries. These pricing models are on-demand, bulk-selling, reserved (two-part tariff) and bulk + reserved. In comparison with the cost-based pricing model, these models allow CSPs to maximize their cloud business revenue and profits. Moreover, CSPs can work with their cloud customers to build a value co-reaction relationship in a service-dominant logic domain.

The experimental results showed that bulk + reserved could increase a CSP's profit margin by 219% in comparison with the cost-based model, while on-demand can lift profit margins by 173%, bulk-selling can quickly achieve a 30% profit margin and lower the sales price by 30%, while the reserved model can reach 213% over the cost-based one.

7.2 Future Directions

In light of the above taxonomy and survey, we predict that cloud pricing is moving further away from a physical box-oriented model to an abstract sandbox-based model. Many CSPs start to offer cloud pricing based on an abstract layer of cloud resources. To some extent, the pricing of the

serverless[26] sandbox can be regarded as modeling No Operation Systems [27] (No OS or NoOps), which is an evolutionary direction from an isolated development environment to an integrated environment of both development and operation or DevOps [28].

However, it does not mean that cloud users can ignore the underlying cloud infrastructure, such as cloud security, workload balancing, horizontal or vertical scaling, auto-failover or high availability, and disaster recovery. All these cloud features will be a part of a CSP's responsibility. They become a part of performance measurements or service-based pricing. Cloud customers do not have to get their hands dirty to tune these cloud features directly. They only need to automate and monitor them and ensure they can be delivered. This is why Kubernetes, Apache Mesos, OpenStack, and Docker [29] Swarm have emerged as essential tools for cloud transformation.

As a result of this transformation, we can observe that each CSP often leverages its business and technology strengths to offer its unique cloud services with innovative pricing models. Based on cloud service delivery models, we argue that AWS can be regarded as online retail-oriented pricing for its IaaS delivery. In other words, "AWS brought the Amazon experience to computing resource delivery." [197] Azure is software application-oriented pricing for its SaaS delivery, and GCP is search engine optimization (SEO) oriented pricing for its PasS. The other CSPs can leverage their own cloud expertise to deliver XaaS, such as e-healthcare, cyber-security, Supervisory Control and Data Acquisition (SCADA), Internet of Things (IoTs) and Business Intelligence Analytics.

Overall, the cloud computing technologies and cloud pricing have four possible development trends: computational resources have moved from statefulness to stateless; IT infrastructure has been transferred from dedicated to a shared base; software development has been gradually shifted from mutability to immutability, and cloud pricing is moving from cost-based to value-based pricing strategy. These trends are leading towards a hyper-converged resource pool for the

[26] Serverless – a cloud computing execution model without a defined server – event driven application deployed model.

[27] NoOps – A programming development approach that allows developers to focus on application development and leave activities of interactions with operation system administrations to a software automation process. It means to take advantage of Platform as a Service (PaaS) to automate application deployment process.

[28] DevOps – means an integrated process to streamline software planning, building, programming, testing, releasing, deploying, operating, monitoring and lifecycle.

[29] Docker – a platform to pack an application with all the dependencies it needs into a single standard unit for the deployment

delivery of cloud services. We can further elaborate on these trends (shown in Figure 7—1) from three perspectives:

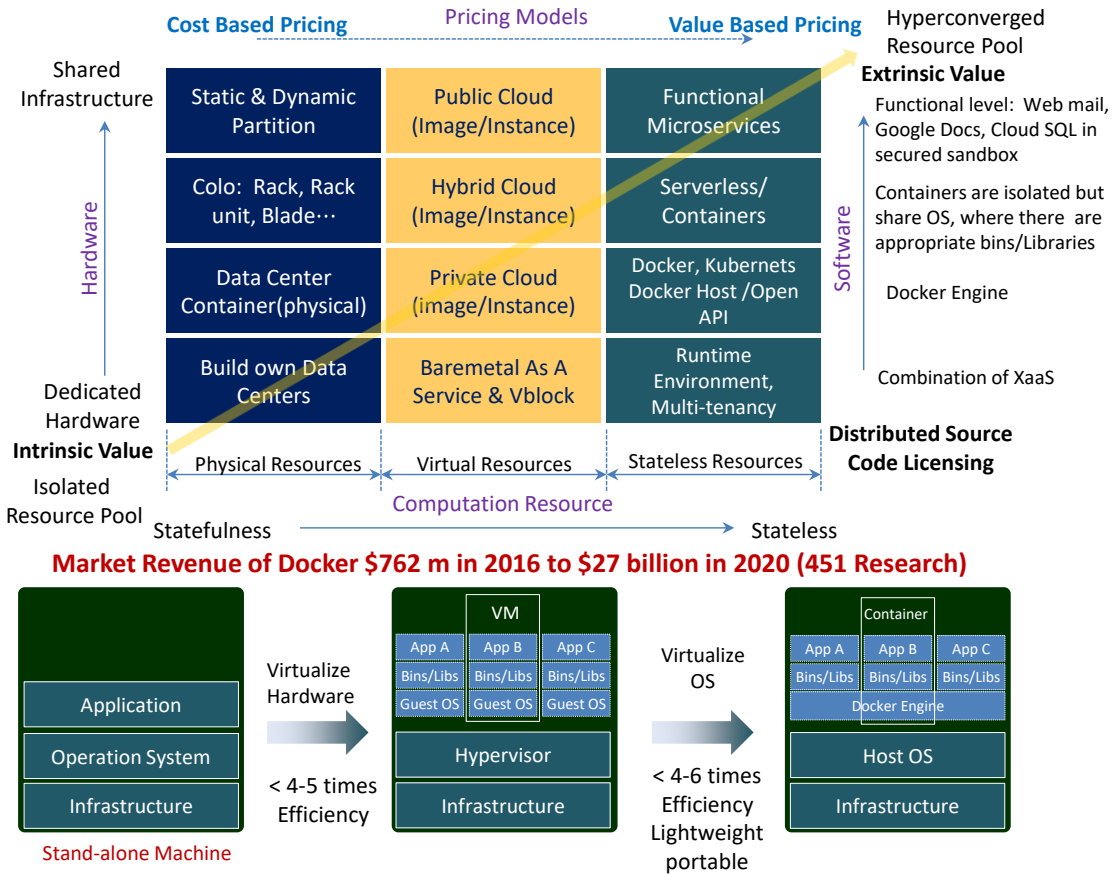


Figure 7—1 Future Trends in Cloud Technologies and Cloud Pricing Strategies [30]

- From an infrastructure or hardware perspective, there is a trend towards sharing, which aims to maximize the utilization of cloud resources. Until now, all cloud infrastructures that are built by either CSPs or large enterprises are supported by the physical data centers and communication networks. If the running business applications require mission-critical infrastructure and satisfy peak demands, the amount of upfront capital and operational expenditures are significant. Yet a large portion (or 90% [60]) of cloud data center capacity might be left idle. Moreover, the value of proportional data center

[30] Microservices – is a software architecture style, which is an evolution of service-oriented architecture (SOA). Function – is a bunch of Lines of Code or sources code to implement a function. Container – an abstraction of operation system (file size in tens of MB Vs VM file size in tens of GB), which is to contain an application only and its environment.

assets is depreciated sharply due to Moore's law. Consequently, sharing infrastructure is an inevitable step to improve the utilization rate of cloud resources. The key difference between cloud resource and traditional infrastructure is that cloud resource is measured by time while the traditional infrastructure refers to a physical object. This leads to a cloud pricing model and its service value being measured by units of time rather than as a physical object.

- From a software perspective, there is a trend towards immutable [31] objects. Traditionally, a software system, such as an operating system, is treated as a mutable object, which is frequently reconfigured and incrementally updated or patched from time to time. For any mutable system, the existing state of software is not cut by one-off but by multiple times on top of older binaries. In comparison, immutable software is a new object. A direct replacement does away with the need for an incremental upgrade. For example, if an old container server needs to upgrade, the fresh new container image will be created and the old one can be thrown away: then the new container is executed. The benefits of immutable software are: 1) the upgrade is traceable if something is going wrong, and 2) it can be rolled back. By moving from mutability to immutability, many software developers can save not only time but also computer resources.
- From a resource perspective, there is a trend towards a stateless architecture that enables customers to scale a resource pool quickly. There are two meanings of stateless. One is a "thin server with the thick client," where a server does not have a memory state of the past and only the client remembers every transaction. Another connotation of stateless is that a workload to be implemented does not need a server that traditionally needs defined memory, network bandwidth, storage, or an operating system. In simple terms, the cloud customer does not need a specified or fixed server box, whether physical or virtual. The workload or application only requires a resource boundary to run or execute functional codes. This temporary resource boundary is also called a container. In comparison with the traditional definition of server resources, the

[31] It is a programming term, which means the value of some objects (e.g. variable, data structure, a function, or a method) can be altered or updated while the term of immutable means the value of the object cannot be changed.

ephemeral nature of the computational requirement can save a lot of cloud resources. This type of computational resource is also considered to be serverless.

All the above three cloud developments not only emphasize the value of hardware but also underscore the value of running business applications. 451 Research estimated that the market revenue of Docker would grow more than 35-fold from $761m in 2016 to $27billion in 2020 [96]. The fundamental reason behind this growth is the efficiency improvement of cloud resources. The initial phase of a cloud transformation from physical to virtual can improve efficiency up to about 4-5 times by reducing cloud infrastructure or cloud data center idle time. The following phase of cloud transformation can increase efficiency by up to 4-6 times by leveraging a server's lightweight (as shown in Figure 7—1)

Figure 7—1 also indicates that Serverless, Docker container, Open API, DevOps, Desktop Grid and Microservices will underpin new cloud pricing models innovation. From a CSP's perspective, the implication of the new cloud technologies allows CSPs to meet the challenge of demand fluctuation and maximize their revenues and profits with a finite amount of cloud resources. From a cloud consumer's perspective, it means cloud vendor-free, flexibility, scalability and Opex minimization. On the basis of this evolutionary trend, we can identify four potential challenges of future cloud price modeling:

- How to drive value-based pricing from a customer's value proposition.
- How to price cloud resources from statefulness to stateless.
- How to price software from mutable to immutable.
- How to price both intrinsic to extrinsic cloud features from a cloud infrastructure lifecycle perspective.

T. Nagle's seminal book [10] provides some clues on how to deal with these challenges. One of the proposals is to establish or consolidate a value-based metric from a cloud business customer's perspective, which is to create a proactive pricing strategy to understand how and when to satisfy the customers' application and meet all its expectations while a CSP can maximize its cloud profit.

7.2.1 Hedonic Pricing for Cloud Computing Services

The conclusion of the hedonic pricing model is that the cloud instance price cannot be just examined by its intrinsic characteristics (mainly cost components, such as RAM, CPU, network performance and storage) alone. It will inevitably lead to a pricing estimation bias because the cloud price prediction is ultimately determined by three key factors or variables, namely, intrinsic, extrinsic and time dummy. Many traditional cloud pricing models cannot reflect cloud extrinsic values (such as burstable CPU, dedicated server, or data center global footprint). However, it does not mean we can ignore these extrinsic characteristics. In fact, they have a substantial influence on the cloud service price. In this thesis, we have shown the process of how to calculate and predict the cloud price accurately and how to avoid a price estimation bias. The novelty of our work is that we present and prove that the value of the Average Annual Growth Rate (AAGR) is equivalent to Moore's law in cloud services.

Chapter 3 argues that the hedonic pricing model is a better approach to estimate the cloud price accurately - if we can establish an adequate hedonic function form based on the available dataset to hand. Furthermore, we exhibit the AWS cloud price has been declining over the last 10 years, but at a much slower pace in comparison with Moore's law prediction. One of the significant influenced factors of this decline is the cloud of extrinsic values or characteristics. They have become AWS's competitive advantages to lead in the cloud (IaaS) market.

Some of the model assumptions can impact the accuracy of cloud price prediction. However, if many CSPs' datasets are opened, this research can improve the prediction results. In the future, any follow-up to this chapter will focus on the combination of both panel and cross-sectional datasets for all leading global CSPs.

7.2.2 Cloud Computing Market Segmentation

Chapter 4 demonstrates how to combine both Hieratical Clustering (HC) and Time Series (TS) forecast to segment the cloud market and predict market demands. In summary, we show HC + TS is a better method to understand the market potential. It is also convenient for any CSP to implement its cloud market strategy by rolling out different pricing models for various market segments. Our approach allows CSPs to tailor their limited cloud resources for the targeted customers. Moreover, CSPs can optimize its cloud pricing beyond the reach of traditional cost-based cloud pricing. It creates opportunities for the CSP to maximize the revenue and profits based on the various cloud customers' utility and surplus. The details of how to define the customer surplus or cloud customer utility functions and how to establish and optimize different

cloud pricing models are among our future work. We will explore these two topics in future studies.

7.2.3 Modeling Cloud Customer Utility Functions

The idea of the modeling cloud utility function is to measure cloud customers' preferences and tastes in terms of less or more VM resources consumed. In our case, the unit of this measurement can be interpreted as a revenue contribution to a cloud business application. There are many factors that can impact overall business revenue, such as responses to time, latency, throughput, cost, SLA guarantee, availability, scalability, capacity and security. These measurements are also called cloud customer metrics. They lead to different utility functions or customer preferences. So, to model multiple utility functions is essential to be successful in cloud pricing.

Chapter 5 demonstrated how to construct different utility functions by the "linked-in" modeling method. We also illustrated how to derive various utility functions in a practical way so that the audiences can build their own utility functions based on different business strategies.

Consequently, Chapter 5 lays out the foundation of cloud price modeling. Our conclusion is that the linked-in modeling method can explicitly model the cloud utility function in each cloud market segment. In comparison with other methods, our modeling method is compelling, comprehensive, flexible and practical for many cloud practitioners.

Practically, the modeling utility function is just one of the process steps for CSPs to create a value-based pricing strategy. There are two more steps: they are to build various value-based price models (Step 3) and identify the optimal price point for each price model so that both CSPs and cloud business customers can achieve the goal of value co-creation (Step 4) as shown in Figure 5—1. In the future, we will complete these two steps of the process based on the results of utility function modeling and provide a comprehensive framework of pricing strategy for CSPs to generate multiple value-based price models from end to end.

7.2.4 Value-Based Cloud Price Modelling For a Segmented B2B Market

Chapter 6 developed an overall framework of the pricing process; that is, how to generate various price models and how to find these optimal price points of each model for a CSP to maximize its profit. These are two elements of pricing strategy (shown in Figure 6—1) that have

been demystified in this paper. The significance of this study is that it presents the complete process of value-based pricing from end to end (E2E).

It demonstrates how to establish four types of practice price models, namely: on-demand, bulk-selling, reserved, and bulk-selling + reserved pricing models. While the modeling process appears to maximize CSP's profit, it is a value co-creation because the process is to generate a partnership between cloud business customers and CSPs. This process becomes a practical tool for any CSP to construct their cloud price models based on the defined business strategy, cloud market environment, own datasets and their expertise.

Chapter 6 shows how to use a Genetic Algorithm (GA) to find the optimal price points by maximizing CSP's profit. Our experimental results demonstrate that the bulk + reserved pricing model can achieve the best profit margin, which is about 219%, while the bulk-selling is the most pervasive model for three market segments in terms of the customers' value propositions. It implies that the single pricing model with an assumption of an integrated market does not necessarily mean it can achieve the maximum profit for CSPs because the cloud market is segmented. Our experimental results reiterate the importance of cloud market segmentation, which has often been ignored by previous studies.

The results also illustrate that the proposed models for most market segments are not able to capture the value of risk-taking (or niche) market segments. The only large discount price model can satisfy the customers who are willing to take high risks for their business application workloads. If a CSP wants to capture the value of a niche market, it should carefully design a particular price model not only to target that niche market segment but also to isolate the model and avoid cannibalizing the higher profit margin from other cloud market segments. On the other hand, the price model that can generate the highest profit margin potential does not necessarily mean that a CSP should adopt it as the only model because there are many CSP competitors on the market. Consequently, our future work will extend from a monopoly market assumption to oligopolies or competitive cloud market environments and from a fixed demand to a price-sensitive demand with a probability distribution.

BIBLIOGRAPHY

[1] J. Weinman, "Cloudonomics: The business value of cloud computing," *John Wiley & Sons*, 2012. p.160

[2] B. Martens, et al., "Costing of cloud computing services: A total cost of ownership approach," *System Science (HICSS), 45th Hawaii International Conference on. IEEE*. 2012

[3] R. Buyya, et al. Cloud computing : principles and paradigms. Hoboken, N.J. : Wiley, c2011

[4] J. Singh and Kumar, V. Multi-Disciplinary Research Issues in Cloud Computing. 2016

[5] G. Pallis, "Cloud Computing: The New Frontier of Internet Computing." IEEE Internet Computing, Internet Computing, no. 5 (2010): 70

[6] M. Engelson, Pricing Strategy : An Interdisciplinary Approach. Joint Management Strategy, 1995

[7] P. Belleflamme, and M. Peitz., "Industrial organization: markets and strategies," *Cambridge University Press*, *New York*, 2015, p. 27

[8] H. R. Varian, "Price discrimination, Handbook of industrial organization," Volume 1, *Elsevier*, 1989, p. 597-654

[9] W. J. Wessels, Economics 3rd Edition, Barron's Education Series Inc. 2000

[10] T. Nagle, et al., The Strategy and Tactics of Pricing : a Guide to Growing More Profitably. 5th ed., International ed. / Boston, Mass.: Pearson, 2011.

[11] Monroe KB, Pricing : Making Profitable Decisions. Boston : McGraw-Hill/Irwin, 2003

[12] https://www.forbes.com/sites/louiscolumbus/2018/09/23/roundup-of-cloud-computing-forecasts-and-market-estimates-2018/#5f62d5b2507b

[13] https://www.gartner.com/en/newsroom/press-releases/2018-09-12-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2019

[14] https://wikibon.com/wikibon-2018-true-private-cloud-forecast-market-shares/

[15] Moore, G.A. and McKenna, R., 1999, Crossing the Chasm, p.5-7

[16] https://tco.vmware.com/tcocalculator/

[17] https://awstcocalculator.com/

[18] https://www.bmc.com/blogs/gartner-magic-quadrant-cloud-iaas/

[19] K. R. Popper, "The open society and its enemies. 5th Edition", *Routledge*, *UK*, 2012, p.68

[20] G. E. Moore, "Principia Ethica," *Dover Publications*, *New York*, 2004, p. 59-108

[21] M. Benioff, and C. Adler, "Behind the cloud: the untold story of how Salesforce.com went from idea to billion-dollar company-and revolutionized an industry," *Jossey-Bass*, *San Francisco*, 2009, p.103-105

[22] A. Iosup and D. Epema, "Grid computing workloads." *IEEE Internet Computing 15.2*, 2011, p.19-26.

[23] W.D. Ross and Philip Stratton-Lake. The Right and the Good. [Electronic Resource]. Oxford : Clarendon Press, 2002, chapter 2

[24] V. A. Zeithaml, "Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence," *The Journal of marketing* 1988, p.2-22.

[25] J. N. Sheth, et al., "Why we buy what we buy: A theory of consumption values," *Journal of business research 22.2*, 1991 p.159-170.

[26] W. Ulaga, and S. Chacour, "Measuring customer-perceived value in business markets: a prerequisite for marketing strategy development and implementation," *Industrial marketing management* 30.6, 2001, p.525-540

[27] A. Eggert, and W. Ulaga, "Customer perceived value: a substitute for satisfaction in business markets?", *Journal of Business & industrial marketing* 17.2/3, 2002, p.107-118

[28] T. F. Bewley, "General equilibrium, overlapping generations models, and optimal growth theory," *Harvard University Press*, US, 2007, p.8 – 16,

[29] E. B. Seufert, “Freemium economics: Leveraging analytics and user segmentation to drive revenue,” Elsevier, 2014, p22

[30] M. F. Niculescu and D. J. Wu. “When should software firms commercialize new products via freemium business models,” *Under Review*, 2011, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.220.9580

[31] T. J. Smith, “Value-Based Pricing, Pricing Done Right: The Pricing Framework Proven Successful by the World's Most Profitable Companies” *Wiley-Blackwell*, 2014, p.11-34.

[32] R. Harmon, et al., “Value-based pricing for new software products: strategy insights for developers,” *the Proceedings of the Portland International Conference on Management of Engineering and Technology*. 2005.

[33] A. E. Boardman, et al., “Cost-benefit analysis: concepts and practice. 4th ed.” *Prentice-Hall, Boston*, 2011, p.27-32

[34] H. M. Cannon, and F. W. Morgan, “A strategic pricing framework,” *Journal of Services Marketing*, 4.2, 1990, p.19-30.

[35] J. L. Forbis, and N. T. Mehta, “Value-based strategies for industrial products,” *Business Horizons*, 24.3, 1981, p.32-42

[36] A. Hinterhuber, “Customer value-based pricing strategies: why companies resist,” *Journal of business strategy*, 29.4, 2008, p.41-50.

[37] W. Elmaghraby, & P. Keskinocak, “Dynamic Pricing in the Presence of Inventory Considerations,” *Research Overview, Current Practices, and Future Directions. Management Science*, 49(10), 2003, p.1287-1309.

[38] A. M. Bandalouski et al. “An Overview of Revenue Management and Dynamic Pricing Models in Hotel Business.” RAIRO-Operation Research, vol. 52, no. 1, pp. 119–141.

[39] N. B. Ruparelia, “Cloud Computing.” *MIT Press. NY, US* 2016, p.17, p.65

[40] B. Shapiro, “Is Performance-Based Pricing the Right Price for You?” 2002 http://hbswk.hbs.edu/item/3021.html

[41] Y. J. Hu, “Performance-based pricing models in online advertising,” *mimeo, MIT Sloan School of Management*, 2004.

[42] http://www.zdnet.com/article/amazon-web-services-marks-40th-price-drop-since-2006/

[43] J. Fernie, et al., “Principles of Retailing,” *Routledge*, 2015 p.370

[44] R. Becerril-Arreola, et al., “Online retailers’ promotional pricing, free-shipping threshold, and inventory decisions: A simulation-based analysis,” *European Journal of Operational Research 230.2*, 2013, p.272-283.

[45] http://revenuesandprofits.com/amazon-web-services-aws-revenues-profits-analysis-2013-2015/

[46] A. Mochón and Y. Sáez, “Understanding Auctions,” *Springer, New York*, p.25.

[47] L. M Ausubel, and J. J. Heckman, “Auction theory for the new economy, New Economy Handbook,” *Elsevier BV, North-Holland*, 2003, p.126-162.

[48] Philip Schofield, ”Jeremy Bentham, the principle of utility, and legal positivism,” *Current Legal Problems 56.1* ,2003, p.1.

[49] R. Buyya(ed.), “High-Performance Cluster Computing: Architectures and Systems, Volume 1 and 2”, *Prentice-Hall, NJ, USA*, 1999, p.9-19

[50] W. H. Bell, et al., “Evaluation of an Economy-Based File Replication Strategy for a Data Grid,” *In International Workshop on Agent-based Cluster and Grid Computing, Tokyo, Japan, IEEE Computer Society Press*, 2003

[51] R. Buyya, et al., “The Grid Economy,” *Proceedings of the IEEE*, 2005, 93(3): p.698-714

[52] R. Buyya, “Market-Oriented Cloud Computing: Vision Hype and Reality of Delivering Computing as the 5th Utility, Cluster Computing and the Grid 2009”. *CCGRID '09. 9th IEEE/ACM International Symposium on*, 2009, p. 1-1.

[53] K. Hwang, et al., “Distributed and Cloud Computing: From Parallel Processing to the Internet of Things,” *Elsevier*. 2013, p.51

[54] R. Buyya, “Economic Models for Resource Management and Scheduling in Grid Computing Concurrency and Computation” *Practice and Experience*, 2002, 14(13-15):p1507-1542

[55] A. N. Toosi, “Revenue Maximization with Optimal Capacity Control in Infrastructure as a Service Cloud Markets,” *IEEE Transaction on Cloud Computing*, 3.3 (2015), p.261-274

[56] H. Xu and B. Li, “Dynamic cloud pricing for revenue maximization,” *IEEE Transactions on Cloud Computing* 1.2, 2013, p.158-171.

[57] G. Gallego, and G. Ryzin, “Optimal dynamic pricing of inventories with stochastic demand over finite horizons,” *Management Science* 40.8, 1994, p.999-1020.

[58] https://aws.amazon.com/blogs/aws/category/ec2-spot-instances/

[59] https://moz.com/devblog/amazon-ec2-spot-request-volatility-hits-1000hour

[60] A. Greenberg, et al., "The cost of a cloud: research problems in data center networks," *ACM SIGCOMM computer communication review* 39.1, 2008, p.68-73.

[61] O. A. Ben-Yehuda, et al., "Deconstructing amazon ec2 spot instance pricing", *ACM Transactions on Economics and Computation* 1.3 (2013): p16.

[62] L. Zheng, et al., "How to bid the cloud," *ACM SIGCOMM Computer Communication Review* 45.4, 2015, p71-84

[63] A. Andrzejak, et al., "Decision model for cloud computing under SLA constraints," *2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, 2010, P.257-266

[64] M. Mazzucco, and M. Dumas, "Achieving performance and availability guarantees with spot instances," *High-Performance Computing and Communications (HPCC), 2011 IEEE 13th International Conference on High-Performance Computing and Communications* , 2011, p.296-303.

[65] Q. Zhang, et al., "Dynamic resource allocation for spot markets in cloud computing environments," *2011 Fourth IEEE International Conference on Utility and Cloud Computing (UCC)*, 2011, p.178-185.

[66] S. Yang, et al., "Optimal Bidding in spot instance market" *INFOCOM, 2012 Proceedings IEEE*. 2012, p.190-198

[67] S.J. Tang, et al., "Towards optimal bidding strategy for Amazon EC2 cloud spot instance*", 2012 IEEE 5th International Conference on Cloud Computing (CLOUD)*, 2012, p.91

[68] S. Yi, et al., "Reducing costs of spot instances via checkpointing in the amazon elastic compute cloud," *2010 IEEE 3rd International Conference on Cloud Computing (CLOUD)*, 2010, p.236-243

[69] J. Chen, et al., "Tradeoffs between profit and customer satisfaction for service provisioning in the cloud," *Proceedings of the 20th international symposium on High performance distributed computing - HPDC '11*, 2011, p.229-238.

[70] W. Gao, et al., "Bidding for highly available services with low price in spot instance market," *Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing - HPDC '15*, 2015, p.191-202

[71] F. C. Schweppe, et al., "Spot pricing of electricity." *Springer US*, 1988, p32

[72] G. Feng, et al., "Revenue maximization using adaptive resource provisioning in cloud computing environments," *2012 ACM/IEEE 13th International Conference on Grid Computing*, 2012, p.192-200,

[73] https://moz.com/blog/crawl-outage

[74] https://blog.serverdensity.com/cloud-vs-colocation/

[75] C. Wu and R. Buyya, "Cloud Data Centers and Cost Modeling: A complete guide to planning, designing and building a cloud data center," *Morgan Kaufmann*, 2015, p.167, p.690

[76] https://www.energycouncil.com.au/analysis/worldwide-electricity-prices-how-does-australia-compare/

[77] E. Walker, "The real cost of a CPU hour," *Computer* 42.4, 2009, p35-41

[78] E. Walter et al., "To lease or not to lease from storage clouds," *Computer* 43.4, 2010, p.44-50.

[79] https://hblok.net/blog/storage/

[80] H. Xu and B. Li., "A study of pricing for cloud resources," *ACM SIGMETRICS Performance Evaluation Review* 40.4 2013, p.3-12.

[81] B. Luderer, et al., "Mathematical formulas for economists," *Springer Science & Business Media, Berlin Heidelberg.* 2009, p.60-89

[82] M. Alam, et al., "Analysis and clustering of the workload in google cluster trace based on resource usage." *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)* ,2015, p.740-747

[83] M. Nakata, "All about RIKEN Integrated Cluster of Clusters (RICC)," *International Journal of Networking and Computing* 2.2, 2012, p.206-215.

[84] S. El Kihal, et al., "Price Comparison for Infrastructure-as-a-Service," *ECIS*, 2012

[85] P. Mitropoulou, et al., "Pricing cloud IaaS services based on a hedonic price index," *Computing* 98.11, 2016, p.1075-1089

[86] L. Zhang, "Price trends for cloud computing services" (2016), https://repository.wellesley.edu/thesiscollection/386/?utm_source=repository.wellesley.edu%2Fthesiscollection%2F386&utm_medium=PDF&utm_campaign=PDFCoverPages

[87] A. Pakes, "A Reconsideration of Hedonic Price Indexes with an Application to PC's.", *American Economic Review* 93.5 (2003): p.1578-1596.

[88] N. Jain, "A truthful mechanism for value-based scheduling in cloud computing," *Theory of Computing* 54.3, 2014, p.388-406.

[89] D. Lucanin, et al., "A cloud controller for performance-based pricing," *2015 IEEE 8th International Conference on cloud computing*, 2015, p.155-162,

[90] A. K. Kar, and A. Rakshit, "Pricing of Cloud IaaS Based on Feature Prioritization-A Value-Based Approach. Recent Advances in Intelligent Informatics". *Springer, Cham*, 2014, p.321-330.

[91] L. Wu, et al., "SLA-based resource allocation for software as a service provider (saas) in cloud computing environments," *2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, 2011, p.195-204

[92] D. Durkee, "Why cloud computing will never be free," *Queue* 8.4, 2010, p.20.

[93] A. Eivy, "Be wary of the economics of "Serverless" Cloud Computing," *IEEE Cloud Computing IEEE*, 4.2, 2017, p.6-12

[94] https://aws.amazon.com/lambda/pricing/

[95] P. Sbarski, "Serverless Architectures on AWS," *Manning Publications NY, US* ,2017, p.2-15

[96] https://451research.com/blog/1351-application-containers-will-be-a-$2-7bn-market-by-2020,-representing-a-small-but-high-growth-segment-of-the-cloud-enabling-technologies-market

[97] H. Xu, and B. Li, 2012, Maximizing revenue with dynamic cloud pricing: The infinite horizon case, 2012 IEEE International Conference on Communications (ICC), ISSN: 1550-3607, Page: 2929-2933

[98] R. J. Dolan, "Pricing: A Value-Based Approach," *Harvard Business School Background* Note 500-071, https://hbr.org/product/Pricing---A-Value-Based-A/an/500071-PDF-ENG

[99] S. Rizou, and A. Polyviou, "Towards value-based resource provisioning in the cloud," *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, 2012, p.155-160

[100] I. Hirose, and J. Olson, "The Oxford Handbook of Value Theory," *Oxford University Press* (OUP), 2015 p.13

[101] https://go.forrester.com/blogs/16-02-02-salesforce_announces_new_pricing_and_packaging_what_it_means_to_you/

[102] J. Gans, et. al. Principles of Economics, Cengage Learning Australia – Melbourne, p1, 978-1-4390-3897-0, 2018

[103] NIST Cloud Computing Service Metrics Description, https://www.nist.gov/publications/cloud-computing-service-metrics-description

[104] Oracle Service Cloud, Customer Experience Metrics and Key Performance Indicators http://www.oracle.com/us/products/applications/cx-metrics-kpi-dictionary-1966465.pdf

[105] https://storageservers.wordpress.com/2012/11/22/microsoft-proves-practically-that-vmware-is-too-expensive/

[106] https://www.dedoimedo.com/computers/vmware-workstation-14.html

[107] Campbell, J., O'Rourke, M., & Slater, M. (Eds.), Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science. MIT Press. 2011, Retrieved from http://www.jstor.org/stable/j.ctt5hhj54

[108] https://iex.ec/

[109] https://www.cloudandheat.com/

[110] https://www.ringcentral.com/

[111] https://databricks.com/

[112] https://cloud.google.com/tpu/docs/pricing

[113] C. Joe-Wong, & S. Sen, Mathematical frameworks for pricing in the cloud: Revenue, fairness, and resource allocations, 2012 arXiv preprint arXiv:1212.0022.

[114] C. Joe-Wong, and S. Sen Pricing the Cloud: Resource Allocations, Fairness, and Revenue. In Workshop on Information Technology & Systems WITS 2013

[115] M. Chiang, et al., Layering as optimization decomposition: A mathematical theory of network architectures. Proceedings of the IEEE, 95(1), 2007, p.255-312.

[116] S. Sen, C. Joe-Wong, S. Ha, S. and M. Chiang, Smart data pricing. John Wiley & Sons., 2014, p.127-166

[117] C. Kilcioglu, and J. Rao, Competition on Price and Quality in Cloud Computing, WWW 2016, April 11–15, Montréal, Québec, Canada ACM, 2016

[118] M. Shahrad, et al., Incentivizing self-capping to increase cloud utilization. In Proceedings of the 2017 Symposium on Cloud Computing, 2017, p. 52-65 ACM.

[119] M. Shahrad, and D. Wentzlaff, Availability knob: Flexible user-defined availability in the cloud. In Proceedings of the Seventh ACM Symposium on Cloud Computing, 2016, October, p. 42-56. ACM.

[120] S. F. Roberts, Measurement Theory with Applications to Decision-making, Utility, and the Social Sciences, Cambridge University Press, 1984, p.6-8

[121] J. L. Daly , Pricing for Profitability: Activity-based Pricing for Competitive Advantage, John Wiley & Sons, Inc. 2002

[122] S. Chen, H. Lee, and K. Moinzadeh, Pricing Schemes in Cloud Computing: Utilization-Based vs. Reservation-Based. Production and Operations Management, 28(1), 2019, pp.82-102.

[123] V.C. Bumgardner., OpenStack in action. Manning Publications Company. 2016, p.5

[124] V. Ramaswamy, and K. Ozcan, What is Co-creation? An Interactional Creation Framework and Its Implications for Value Creation, Journal of Business Research, vol. 84, 2017, p. 196-205

[125] R. Buyya, et al. (ed.), Cloud computing: Principles and paradigms, Vol. 87. *John Wiley & Sons*, 2010.

[126] L. Venkatachalam, "The contingent valuation method: a review," *Environmental impact assessment review* 24.1, 2004, p. 89-124.

[127] H. A. Linstone, and T. Murray, "The Delphi method Techniques and Applications, Vol. 29". *Reading, MA: Addison-Wesley*, 1975.

[128] K. Lampe, "The Birth of Hedonism: The Cyrenaic Philosophers and Pleasure as a Way of Life," *Princeton University Press*, 2014.

[129] J. Bentham, "The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation," *Clarendon Press*, 1996.

[130] J. S. Mill, "Utilitarianism, Chapter II, What Utilitarianism Is, Chapter II, Of the Ultimate Sanction of the Principle of Utility"' 1863.

[131] T. Hurka, "Moore in the Middle," *Ethics* 113.3, 2003, p. 599-628.

[132] Z. Griliches, "Hedonic price indexes for automobiles: An econometric of quality change," *The Price Statistics of the Federal Government. NBER*, 1961, p. 173-196.

[133] R. Michaels, "Hedonic prices and the structure of the digital computer industry," *the Journal of Industrial Economics* 1979, p. 263-275.

[134] R. Cole, et al., "Quality-adjusted price indexes for computer processors and selected peripheral equipment," *Survey of Current Business* 66.1, 1986, p. 41-50.

[135] E.R. Berndt, and Z. Griliches, "Price indexes for microcomputers: an exploratory study," *Price measurements and their uses*, *University of Chicago Press*, 1993, p. 63-100.

[136] R. H. Raghav, and B. D. Lynch, "Hedonic price analysis of workstation attributes," *Communications of the ACM* 36.12 1993, p. 95-102.

[137] R.M. Sakia, "The Box-Cox transformation technique: a review," *The statistician* 1992, p. 169-178.

[138] A. Pakes, "Reconsideration of Hedonic Price Indexes with an Application to PC's," *The American Economic Review* 93.5 2003, p. 1578-1596.

[139] C. R. Hulten, "Price hedonics: a critical review," *Economic Policy Review*, Vol.9 No.3, 2003. https://ssrn.com/abstract=788904

[140] P. Davis, and G. Eliana, "Quantitative techniques for competition and antitrust analysis," *Princeton University Press*, 2009, p.286-287

[141] K.J. Lancaster, "A new approach to consumer theory," *Journal of political economy* 74.2, 1966, p. 132-157.

[142] S. Rosen, "Hedonic prices and implicit markets: product differentiation in pure competition, " *Journal of political economy* 82.1, 1974, p. 34-55.

[143] H. W. Brachinger, "Statistical theory of hedonic price indices," *University de Fribourg*, 2002.

[144] Internet Archive, https://archive.org/web/

[145] AWS annual report, http://phx.corporateir.net/phoenix.zhtml?c=97664&p=irol-reportsannual/

[146] R. Halvorsen, and H. O. Pollakowski, "Choice of functional form for hedonic price equations," *Journal of urban economics* 10.1, 1981, p. 37-49.

[147] E. Cassel, and R. Mendelsohn, "The choice of functional forms for hedonic price equations: comment." *Journal of Urban Economics* 18.2, 1985, p.135-142.

[148] J. E. Triplett, "Hedonic methods in statistical agency environments: an intellectual biopsy." *Fifty years of economic measurement: the jubilee of the conference on research in income and wealth. University of Chicago Press*, 1991.

[149] J.E. Triplett, "Draft copy Handbook on a quality adjustment of price indexes for information and communication technology products," *Paris: OECD*, 2000.

[150] Z. Griliches, "Price indexes and quality change," *Harvard University Press*, 1971.

[151] Z. Griliches, "Hedonic price indexes and the measurement of capital and productivity: some historical reflections." *Fifty years of economic measurement: The Jubilee of the conference on research in income and wealth, University of Chicago Press*, 1991.

[152] J. R. Gordon, "The measurement of durable goods prices," *University of Chicago Press*, 1990, p.188-240.

[153] J. H. McDonald, "Handbook of biological statistics, Vol. 2", Baltimore, MD: *Sparky House Publishing*, 2009.

[154] N.V. Kuminoff, et al., "Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities?" *Journal of Environmental Economics and Management* 60.3 2010, p.145-160.

[155] M. L. Cropper, et al., "On the choice of functional form for hedonic price functions," *The Review of Economics and Statistics* 1988, p. 668-675.

[156] H. Kyle, "AWS moves from ECU to vCPU," http://blogs.gartner .com/kylehilgendorf/2014/04/16/aws-moves-from-ecu-to-vcpu/

[157] J. Read, "What is an ECU? CPU Benchmarking in the Cloud", http://blog.cloudharmony.com/20-10/05/what-is-ecu-cpu-benchmarking-in-cloud.html

[158] C. Mack, "The Multiple Lives of Moore's Law," http://spectrum .ieee.org/magazine/2015/April

[159] M. Silver, and S. Heravi, "The Difference between Hedonic Imputation Indexes and Time Dummy Hedonic Indexes," *Washington, D.C. International Monetary Fund*, 2006.

[160] A. Aizcorbe, "The Stability Of Dummy Variable Price Measures Obtained From Hedonic Regressions," *Mimeo, Federal, Reserve Board, Washington DC*, 2003

[161] M. Naldi, and L. Mastroeni, "Economic decision criteria for the migration to cloud storage," *European Journal of Information Systems*, 25(1): 2016, p.16-28

[162] P. Mitropoulou, et al., "Pricing IaaS: A Hedonic Price Index Approach," *Lecture Notes in Computer Science* 2017, p.18-28

[163] A. Baranzini, et al., "Hedonic methods in housing markets: Pricing environmental amenities and segregation," *Springer Science & Business Media*, 2008

[164] Storage, Full history 1957 –present, https://hblok.net/blog/storage/

[165] W. Wang, et al., "Towards Optimal Capacity Segmentation with Hybrid Cloud Pricing." 2012 IEEE 32Nd International Conference On Distributed Computing Systems, Distributed Computing Systems (ICDCS), 2012, p425

[166] S. Berman, et al., "How cloud computing enables process and business model innovation," *Strategy & Leadership*.40(4), 2012 p27-35

[167] http://www.businessinsider.com/visa-partners-with-banks-for-cross-border-b2b-payments-2017-11/?r=AU&IR=T

[168] R. Smith, "Product Differentiation and Market Segmentation as Alternative Marketing Strategies." *Journal of Marketing, vol. 21*, no. 1, 1956, p.3–8.

[169] D. Yankelovich, and D. Meer, "Rediscovering Market Segmentation," *Harvard Business Review*, Vol. 84, Issue 2, 2006, p.122-131

[170] J. Claycamp, and M. William, "A Theory of Market Segmentation." *Journal of Marketing Research*, vol. 5, no. 4, 1968, p. 388–394.

[171] M. Wedel, & W.A. Kamakura, "Market segmentation: conceptual and methodological foundations," *Boston: Kluwer Academic*, 1998

[172] M. McDonald, and I. Dunbar, "Market Segmentation. how to do it and how to profit from it" , *John Wiley & Sons*. 2012

[173] Y. Wind, and R. Cardozo, "Industrial Market Segmentation," *Industrial Marketing Management*, vol. 3, no. 3, 1974, p. 153-165.

[174] P. Kotler, "Marketing Mix Decisions for New Products," *Journal Of Marketing Research (JMR)* 1(1): 1964, p. 43-49.

[175] D. Abramson et al., "A computational economy for grid computing and its implementation in the Nimrod-G resource broker," *Future Generation Computer Systems*. 1; 2002, p18

[176] B. Javadi, et al., "Statistical Modeling of Spot Instance Prices in Public Cloud Environment," *2011 Fourth IEEE International Conference On Utility And Cloud Computing, Utility And Cloud Computing (UCC)*, 2011, p219

[177] P. Hande, et al., "Pricing under Constraints in Access Networks: Revenue Maximization and Congestion Management," *Proceedings IEEE INFOCOM*, 2010

[178] https://github.com/google/cluster-data/blob/master/TraceVersion1.md

[179] https://hbr.org/2005/09/building-loyalty-in-business-markets

[180] L. Wang, and K. Ramamohanarao, "iVAT, and aVAT: Enhanced Visual Analysis for Cluster Tendency Assessment," *Lecture Notes in Computer Science*:16, 2010

[181] R.G. Lawson, and P.C. Jurs, "New index for clustering tendency and its application to chemical problems," *Journal of Chemical Information & Computer Sciences*., Vol. 30 Issue 1, 1990, p36-41.

[182] H.J. Brain, "A New Method for Determining the Type of Distribution of Plant Individuals," *Annals Of Botany*;(70):1954, p.213

[183] https://www.decisionanalyst.com/whitepapers/marketsegmentation/

[184] L. Wang, and K. Ramamohanarao, "Automatically Determining the Number of Clusters in Unlabeled Data Sets," *IEEE Transactions On Knowledge And Data Engineering*, 3), 2009, p.335.

[185] https://www.jstatsoft.org/article/view/v061i06

[186] B. Shapiro, and T. Bonoma, "How to segment industrial markets," *Harvard Business Review*.62(3): 1984, pp.104-110

[187] R. Shumway, and D. Stoffer, "Time Series Analysis And Its Applications: With R Examples," *New York: Springer*, 2011

[188] R. Kuo, et al., "An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation," *Neurocomputing September* 12; 205: 2016, p.116-129

[189] Y. Wind, et al., "Market-based Guideline for Design of Industrial Products," *Journal Of Marketing*. ;42(3): 1978, p. 27-37.

[190] T. Verhallen, et al., "Strategy-Based Segmentation of Industrial Markets," *Industrial Marketing Management*; 27(4): 1998, p.305-313.

[191] R. Best, "An Experiment in Delphi Estimation in Marketing Decision Making," *Journal Of Marketing Research (JMR)* November;11(4): 1974, p. 448-452

[192] S. Freemium, "Economics: Leveraging Analytics And User Segmentation To Drive Revenue," *Waltham, Massachusetts: Morgan Kaufmann*. 2014

[193] J. Laughlin, and C. Taylor, "An approach to industrial market segmentation," *Industrial Marketing Management*, 1991 p.127

[194] R.A. Plank, "Critical Review of Industrial Market Segmentation" *Industrial Marketing Management*; 14(2): 1985 p79-91

[195] G. Balakrishna, "Better Used the industrial Marketing Concept," *Industrial Marketing Management*, 9(1): 1980, p71-76

[196] global e-commerce, 2017 https://www.shopify.com/enterprise/global-ecommerce-statistics

[197] https://www.nist.gov/news-events/news/2011/10/final-version-nist-cloud-computing-definition-published

[198] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," II. *Wiley Interdisciplinary Reviews-Data Mining And Knowledge Discovery*, 2 1, 2012, p86-p97, 12p.

[199] StatCounter, 2018, https://en.wikipedia.org/wiki/Usage_share_of_operating_systems

[200] K. Shin, 'The Executor of Integrated Marketing Communications Strategy": *Marcom Manager's Working Model'* 2013

[201] R.A. Oliva, “a High-Level overview: a value perspective on the price of Business to Business market, Handbook of B2B Marketing”, *Cheltenham, Gloucestershire*: *Edward Elgar Publishing*, 2012, pp.15-40

[202] R.J. Thomas, “B2B market segmentation, Handbook of B2B marketing”, *Cheltenham, Glouces -tershire: Edward Elgar Publishing*, 2012, pp.182-207

[203] P.W. Bridgman, “The Logic of Modern Physics,” *New York : The Macmillan Company*, 1927

[204] B. Fine, et al., "The Driving Force of the Market” *Essays in Austrian Economics*. 2001: p57

[205] https://www.nist.gov/sites/default/files/documents/itl/cloud/RATAX-CloudServiceMetricsDescription-DRAFT-20141111.pdf

[206] R. P. Doyle, et al., 'Model-Based Resource Provisioning in a Web Service Utility,' *Proceedings of the Usenix Symposium Technologies and Systems*, 2003, p. 57

[207] K. Appleby, et al. “Oceano-SLA based management of a computing utility,” 2001 IEEE/IFIP International Symposium on Integrated Network Management Proceedings. Integrated Network Management VII. Integrated Management Strategies for the New Millennium (Cat. No.01EX470), Integrated Network Management Proceedings, 2001 IEEE/IFIP International Symposium on, Integrated network management, 2001, p.855

[208] W.E. Walsh, et al., “Utility functions in autonomic systems,” 2004, International Conference on Autonomic Computing, Proceedings, Autonomic Computing, 2004. Proceedings. International Conference on, Autonomic computing, 2004, p.70

[209] M.N. Bennani, and D.A. Menasce, “Resource Allocation for Autonomic Data Centers using Analytic Performance Models,” 2005, Second International Conference on Autonomic Computing (ICAC'05), Autonomic Computing, ICAC 2005. Proceedings. Second International Conference on, 2005, p. 229

[210] J.O. Kephart, and R. Das, “Achieving Self-Management via Utility Functions,” IEEE Internet Computing, Internet Computing, IEEE, IEEE Internet Computer, no. 1. Available from: 10.1109/MIC. 2007

[211] I. Menache, et al., “Socially optimal pricing of cloud computing resources,” *In VALUE TOOLS* 2011.

[212] E. Weintraub and Y. Cohen, “Optimizing User's Utility from Cloud Computing Services in a Networked Environment,” *International Journal of Advanced Computer Science and Applications*, Vol. 6, No. 10, 2015, p153-163.

[213] D. Burda, and F. Teuteberg, “Exploring consumer preferences in cloud archiving – a student's perspective,” *Behavior & Information Technology,* vol. 35, no. 2, 2016, p. 89-105

[214] D. Minarolli and B. Freisleben, “Utility-based resource allocation for virtual machines in Cloud computing,” 2011 IEEE Symposium on Computers and Communications (ISCC), Computers and Communications (ISCC), 2011 IEEE Symposium on, 410. 2011

[215] M. Cardosa, et al., 'Shares and utilities based power consolidation in virtualized server environments' 2009, 2009 IFIP/IEEE International Symposium On Integrated Network Management, Integrated Network Management, 2009. IM '09. IFIP/IEEE International Symposium On, 2009, p.327

[216] S. Garg, et al., 'SLA-based virtual machine management for heterogeneous workloads in a cloud datacenter,' *Journal Of Network And Computer Applications*, 45, 2014, p. 108-120

[217] J. Yu, R. Buyya, and K. Ramamohanarao, 'Workflow Scheduling Algorithms for Grid Computing,' *Metaheuristics for Scheduling in Distributed Computing Environments*, 2008, p.173

[218] F.S. Roberts, 'Encyclopedia of Mathematics and Its Applications. Volume 7. Measurement Theory with Applications to Decision-making, Utility and the Social Sciences Fred S. Roberts', *Current Science*, no. 13, 2009, p.6-8

[219] W.J. Stewart, “Probability, Markov chains, queues, and simulation : the mathematical basis of performance modeling,” *Princeton, N.J. ; Oxford : Princeton University Press*, 2011, p.193

[220] A. Undheim, et al. 'Differentiated Availability in Cloud Computing SLAs,' 2011 IEEE/ACM 12th International Conference on Grid Computing, Grid Computing (GRID), 2011 12th IEEE/ACM International Conference on, 2011, p. 129

[221] U. Bhat "An Introduction To Queueing Theory : Modeling And Analysis In Applications," *Boston : Birkhäuser*, p.34-40,

[222] C. Muris, et al., 'Pareto utility,' *Theory And Decision, 75*, 2013 1, p. 43-57

[223] http://www.abs.gov.au/

[224] P.R. Krugman, and R. Wells, Economics Fourth Edition, *New York : Worth Publishers*, 2015, p.282-283

[225] S. Landsburg, "Price Theory And Applications" *Minneapolis/St. Paul West Pub. Co*., c1995

[226] G. Box, "An Accidental Statistician : The Life And Memories Of George E.P. Box / George E.P. Box," Hoboken, New Jersey : *John Wiley and Sons*, Inc., 2013

[227] https://docs.oracle.com/en/cloud/iaas/compute-iaas-cloud/stcsg/viewing-instance-metrics.html

[228] H. A. Schmid, G. Rossi, Modeling, and Designing Processes in E-Commerce Applications." IEEE Internet Computing, Internet Computing, no. 1, 2004, p. 19

[229] N. Ranaldo, et al. "Capacity-Aware Utility Function for SLA Negotiation of Cloud Services," 2013 IEEE/ACM 6Th International Conference On Utility And Cloud Computing, Utility And Cloud Computing (UCC), 2013 IEEE/ACM 6Th International Conference On, Utility And Cloud Computing, IEEE International Conference On. 2013, p.292

[230] C. Kilcioglu and J. M. Rao. Competition on price and quality in cloud computing, https://dl.acm.org/citation.cfm?id=2883043

[231] R. Pal, and P. Hui, "Economic models for cloud service markets: Pricing and Capacity planning," *Theoretical Computer Science*, *(Distributed Computing and Networking (ICDCN 2012),* 2013 p.113-124

[232] G. Baltas, and P. Doyle, "Random utility models in marketing research: a survey," *Journal of Business Research*, 2001,p115

[233] M. Young, "Implementing Cloud Design Patterns For AWS : Create Highly Efficient Design Patterns For Scalability, Redundancy, And High Availability In The AWS Cloud," *Birmingham, Packt Publishing*, 2015

[234] J. Schaffner, "Multi-tenancy for cloud-based in-memory column databases workload management and data placement," *Springer*, 2014

[235] http://downtimecost.com/

[236] https://www.microsoft.com/enau/download/confirmation.aspx?id=42026

[237] https://www.asbfeo.gov.au/sites/default/files/Small_Business_Statistical_Report-Final.pdf Accessed in 20/Aug/2018

[238] Marcos-Cuevas, J, et al. "Value Co-Creation Practices and Capabilities: Sustained Purposeful Engagement across B2B Systems." Industrial Marketing Management, vol. 56 July 2016, pp. 97–107

[239] Alves, H., et al. "Value Co-Creation: Concept and Contexts of Application and Study." Journal of Business Research, vol. 69 May 2016, pp. 1626–1633

[240] K. Maqbool, et al. "OSaaS: Online Shopping as a Service to Escalate E-Commerce in Developing Countries." 2016 IEEE 18th International Conference on High-Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2016, p. 1402

[241] 2017 Top 10 European Cloud Providers, Cloud Spectator. https://cloudspectator.com/top-10-european-cloud-service-providers/

[242] M. Koehler, S. Benkner, "Design of an Adaptive Framework for Utility-Based Optimization of Scientific Applications in the Cloud." 2012 IEEE Fifth International Conference on Utility and Cloud Computing, Utility and Cloud Computing (UCC), 2012 IEEE Fifth International Conference on, Utility and Cloud Computing, IEEE International Conference On, 2012, p. 303

[243] Wu, C. Buyya, R. Ramamohanarao K, Value-Based Cloud Pricing Modeling for Maximizing Profit, 2018.

[244] V. Datla , K. Goseva-Popstojanova , Measurement-Based Performance Analysis of e-Commerce Applications with Web Services Components." IEEE International Conference on E-Business Engineering (ICEBE'05), e-Business Engineering, 2005, p. 30

[245] J.D. Strunk, et al. “Using Utility to Provision Storage Systems.” Proceedings of the fast conference on file and storage technologies, 2008, p. 313

[246] P.J. Thomas, “Measuring Risk-Aversion: The Challenge.” Measurement, vol. 79, Feb. 2016, pp. 285–301.

[247] H. Morshedlou, and M. R. Meybodi, Decreasing Impact of SLA Violations:A, Proactive Resource Allocation Approach for Cloud Computing Environments.” IEEE Transactions on Cloud Computing, no. 2, 2014, p. 156

[248] A. Kim, and I.S. Moskowitz, Incentivized Cloud Computing: A Principal-Agent Solution to the Cloud Computing Dilemma. 2010. http://oai.dtic.mil/oai/oai?&verb=getRecord&metadataPrefix=html&identifier=ADA530441

[249] J. Chen, et al. “Tradeoffs Between Profit and Customer Satisfaction for Service Provisioning in the Cloud.” Proceedings of the 20th International Symposium: High-Performance Distributed Computing, June 2011, p. 229

[250] M. Macías, and J. Guitart, “A genetic model for pricing in cloud computing markets,” *Proceedings of the 2011 ACM Symposium: Applied Computing*, 2011, p.113

[251] C. Kilcioglu, and J. Rao, “Competition on Price and Quality in Cloud Computing,” *Proceedings of the 25th International Conference on World Wide Web,* ACM, 2016,

[252] M. Aazam, et al., “Cloud Customer's Historical Record Based Resource Pricing.” *IEEE Transactions On Parallel And Distributed Systems*. n.d.;27(7): 2016, p.1929-1940

[253] C.S. Yeo, et al., “Autonomic metered pricing for a utility computing service,” *Future Generation Computer Systems* January 1, 2010;26 p.1368-1380.

[254] L. Du, “Pricing and Resource Allocation in a Cloud Computing Market,” 12Th IEEE/ACM International Symposium On Cluster, Cloud And Grid Computing (Ccgrid 2012), Cluster, Cloud And Grid Computing (Ccgrid). 2012, p. 817.

[255] J. Sherwani, et al., 'Libra: a computational economy-based job scheduling system for clusters”, *software practice and experience*, no. 6, 2004, p. 573

[256] C.S. Yeo, and R. Buyya, 'Pricing for utility-driven resource management and allocation in clusters,' *International Journal of High-Performance Computing Applications*, no. 4, 2007, p. 405.

[257] P. J. Davis, E. Garcés, “Quantitative techniques for competition and antitrust analysis,” *Princeton : Princeton University Press*, 2010.

[258] https://www.rightscale.com/blog/cloud-cost-analysis/aws-vs-azure-vs-google-cloud-pricing-compute-instances, Accessed 30/Aug/2018

[259] A. Eiben, and J. Smith, “Introduction To Evolutionary Computing,” *Berlin: Springer*, 2015

[260] U. Bhat, “An Introduction To Queueing Theory: Modeling And Analysis In Applications,” *Boston : Birkhäuser*, 2015, p.34-40

[261] J. Norstad, “An Introduction To Utility Theory,” 1999, http://seshadri.us/docs/norstad_utility.pdf Accessed in 15/Aug/2018

[262] A. Undheim, et al., “Differentiated Availability in Cloud Computing SLAs,” *2011 IEEE/ACM 12th International Conference on Grid Computing, on*, 2011, p. 129

[263] M. Zhang, “Utility Functions in Autonomic Workload management for DBMSs,” *International Journal on Advances In Intelligent Systems*, Vol5, No 1&2, 2012

[264] A. Homer, et al., “Cloud Design Patterns: Prescriptive Architecture Guidance for Cloud Applications,” *Microsoft patterns & practices*, 2014. p. 150

[265] S.N. Sivanandam, S. N. Deepa, “Introduction to genetic algorithms,” *New York : Springer*, 2007

[266] R. Leardi, “Nature Inspired Methods In Chemometrics: Genetic Algorithms And Artificial Neural Networks,” *Elsevier*; 2003

[267] https://github.com/google/cluster-data/blob/master/TraceVersion1.md Accessed in 20/Aug/2018

[268] https://www.asbfeo.gov.au/sites/default/files/Small_Business_Statistical_Report-Final.pdf Accessed in 20/Aug/2018

[269] I. Hirose, J. Olson, “ The Oxford Handbook of Value Theory,” *Oxford University Press* (OUP), 2015, p.13

[270] C. Prahalad, V. Ramaswamy, "Co-opting Customer Competence," *Harvard Business Review*, January 2000;78(1):p79-87

[271] V. Ramaswamy, and K. Ozcan, 'What is co-creation? An interactional creation framework and its implications for value creation', *Journal Of Business Research*, 2017, vol. 84, p. 196-205.

[272] M. Kohtamaki, and R. Rajala, 'Theory and practice of value co-creation in B2B systems', *Industrial Marketing Management*, 2016, vol. 56, p. 4-13

[273] R. Lusch, and S. Vargo, 'Gaining competitive advantage with service-dominant logic,' Handbook on business-to-business marketing, 2012, *Edward Elgar*, p. 109-124

[274] C. Grönroos, "Quo Vadis, Marketing? Toward a Relationship Marketing Paradigm", *Journal of Marketing Management, vol. 10, no.* 5, 1994, p. 347-360

[275] F. Alzhouri., A. Agarwal, 'Dynamic Pricing Scheme: Towards Cloud Revenue Maximization,' IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom), Cloud Computing Technology and Science , 2013 IEEE 5th International Conference on, 2015 p. 168

[276] https://www.amd.com/Documents/AMD_WP_Virtualizing_Server_Workloads-PID.pdf Accessed in 20/Aug/2018

[277] V. Tarasov, et al. "Virtual machine workloads: the case for new benchmarks for NAS." *Fast*. 2013, p307-320

[278] P. Hande, et al. "Pricing under constraints in access networks: Revenue maximization and congestion management," *In INFOCOM, 2010 Proceedings IEEE* 2010, p.1-9

[279] J. Mo, and J. Walrand, "Fair end-to-end window-based congestion control." *IEEE/ACM Transactions on Networking*, 8(5), 2000, p.556-567

[280] P. Bats, "Scalability and Economics of Citrix XenApp and Citrix Xen Desktop 7.6 on Amazon Web Services", 2014, https://www.citrix.com/content/dam/citrix/en_us/documents/partner-documents/scalability-and-economics-of-citrix-xenapp-and-citrix-xendesktop-7.6-on-amazon-web-services.pdf Accessed 28/Aug/18

[281] F. Schimscheimer, "Workload Considerations for Virtual Desktop Reference Architecture," https://www.vmware.com/techpapers/2009/workload-considerations-for-virtual-desktop-refer-10081.html , Accessed 28/Aug/2018

[282] P.H. Nakhai, and N.B. Anuar, "Performance evaluation of virtual desktop operating systems in virtual desktop infrastructure," 2017 IEEE Conference on Application, Information and Network Security (AINS), Application, Information and Network Security (AINS), 2017 IEEE Conference on, 2017, p.105

[283] P.P. Wakker, 'Explaining the characteristics of the power (CRRA) utility family,' *Health Economics –Chichester-*, 12, 2008, p.1329

[284] M. Fahrioglu, and F.L. Alvarado, 'Using utility information to calibrate customer demand management behavior models,' *IEEE Transactions on Power Systems, Power Systems, IEEE Transactions on, IEEE Trans. Power Syst, no*. 2,2001, p. 317

[285] A. Mishra, et al., 'Towards Characterizing Cloud Backend Workloads: Insights from Google Compute Clusters,' *Performance Evaluation Review*, 2010, 4, p. 34,

[286] A. Williams et al., Web Workload Characterization: Ten Years Later. In:

[287] X. Tang , J. Xu, Chanson S.T. (eds) Web Content Delivery. Web Information Systems Engineering and Internet Technologies Book Series, vol 2. 2005, *Springer, Boston*, MA.

[288] M.C. Calzarossa, et al., "Workload characterization: A survey revisited," *ACM Computing Surveys (CSUR)*, 2016, 48(3), p.48.

[289] T.W. Manikas, and J.T. Cain, "Genetic algorithms vs. simulated annealing: A comparison of approaches for solving the circuit partitioning problem," 1996, https://pdfs.semanticscholar.org/d7ac/71ec88fc9e5a63f44b950e32d65eaf3b1c2f.pdf Accessed 30/Aug/2018 Accessed in 28/Aug/2018

[290] J.H. Holland, “Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence,” *Cambridge, Massachusetts : The MIT Press*, 1992

[291] M. Maurer, et al., “Cost-benefit analysis of an SLA mapping approach for defining standardized Cloud computing goods,” *Future Generation Computer Systems*, 28(1), 2012, p.39-47.

[292] S. Basu, et al. “Pricing cloud services—the impact of broadband quality,” *Omega,* 50, 2015, p.96-114.

[293] J. Altmann, and M.M. Kashef, “Cost model-based service placement in federated hybrid clouds” *Future Generation Computer Systems*, 41, 2014, p.79-90.

[294] B. Maguire, "The value-based theory of reasons." *The University of North Carolina at Chapel Hill* 2016,

[295] https://www.slideshare.net/AmazonWebServices/retiring-technical-debt-aws-partner-summit-mumbai-2018pdf

[296] J. K. Campbell, et al. “Carving Nature at Its Joints : Natural Kinds in Metaphysics and Science,” *Cambridge, Massachusetts : The MIT Press*, 2011

[297] M. Böhm, S. Leimeister, C. Riedl, and H. Krcmar, Cloud computing–outsourcing 2.0 or a new business model for IT provisioning?. In Application management 2011, p. 31-56. Gabler.

[298] A. Rosenthal, P. Mork, M. H. Li, J. Stanford, D. Koester, and P. Reynolds, Cloud computing: a new business paradigm for biomedical information sharing. Journal of biomedical informatics, 43(2), 2010, p342-353.

[299] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi, Cloud computing—The business perspective. Decision support systems, 51(1), 2011, p176-189.

[300] A. Fox, R. Griffith,A. Joseph, R. Katz, A. Konwinski, G. Lee, G., and I. Stoica, Above the clouds: A Berkeley view of cloud computing. Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS, 28(13), 2009.

[301] C. Weinhardt, A. Anandasivam, B. Blau, N. Borissov, T. Meinl, W. Michalk, and J. Stößer, Cloud computing–a classification, business models, and research directions. Business & Information Systems Engineering, 1(5), 2009, p391-399

[302] J. Altmann, et al. Grid Economics and Business Models : 6th International Workshop, GECON 2009, Delft, the Netherlands, August 24, 2009 : Proceedings. Lecture Notes in Computer Science: 5745. Berlin ; New York : Springer, 2009

[303] J. De Koning, Service Design Geographies. Proceedings of the ServDes. 2016 Conference. No. 125. Linköping University Electronic Press, 2016

[304] S. Novani, Value Co-Creation on Cloud Computing: A Case Study of Bandung City, Indonesia, Systems Science for Complex Policy Making: A Study of Indonesia, edited by Kuntoro Mangkusubroto et al., Translation Systems Sciences, vol. 3. New York: Springer Nature, 2016: p. 43–63

[305] E. Brynjolfsson, et al., Cloud Computing and Electricity: Beyond the Utility Model. Communications of the ACM, 53(5), 2010: p.32-34.

[306] I. A. Kash and P.B. Key, Pricing the Cloud, IEEE Internet Computing, Internet Computing no. 1, 2016: p. 36-43

[307] D. Poola, K. Ramamohanarao, and R. Buyya, Enhancing the Reliability of Workflow Execution Using Task Replication and Spot Instances, ACM Transaction on Autonomous and Adaptive Systems, Vol 10. No.4, 2016: p.1-21

[308] https://www.census.gov/

[309] B. Rady, Serverless Single Page Apps, Fast, Scalable and Available, Pragmatic Bookshelf; 1 edition, June 14, 2016: p3-4

[310] O. C. Ibe, Markov Processes for Stochastic Modeling, Amsterdam, Netherlands : Elsevier, 2013, p55-82

[311] F. Schimscheimer, Workload Considerations for Virtual Desktop Reference Architecture, https://www.vmware.com/techpapers/2009/workload-considerations-for-virtual-desktop-refer-10081.html , Accessed 28/Aug/2018

[312] S. Rylander, B. Gotshall, "Optimal Population Size and the Genetic Algorithm." Population 100.400, 2002: 900

[313] R. C. Picker, The Razors-and-Blades Myth(S), University of Chicago Law Review, no. Issue 1, 2011: p. 225

[314] O. Michalski, and S. Demiliani Implementing Azure Cloud Design Patterns, Birmingham, Packt Publishing, 2018: p109-119

[315] G. D. Feitelson, Workload modeling for performance evaluation, IFIP International Symposium on Computer Performance Modeling, Measurement, and Evaluation. Springer, Berlin, Heidelberg, 2002

[316] J. Luetkehoelter, What Is Disaster Recovery?, Pro SQL Server Disaster Recovery (2008): p.1-12.

[317] https://www.citrix.com.au/global-partners/amazon-web-services/xendesktop-on-aws.html

[318] A. Paul, Citrix XenApp 7.5 Desktop Virtualization Solutions, Plan, Design, Optimize, and Implement Your XenApp Solution to Mobilize Your Business. Birmingham, Packt Publishing, 2014

[319] Citrix Virtual Apps and Desktops on AWS, Citrix, 2014 https://www.citrix.com.au/global-partners/amazon-web-services/xendesktop-on-aws.html

[320] N. Gregory. Mankiw, Principles of economics. Cengage Learning, 2014: p.425, p.447

[321] G. A. Jehle, and P. J. Reny, Advanced microeconomic theory, Pearson Education Edinburgh Gate, Harlow, 2011: p.4, p288

[322] JN. Franklin, Methods of mathematical economics: linear and nonlinear programming, fixed-point theorems. SIAM; 2002: p. 190

[323] K. Matthias, S. Thomas, and S. Horst, Variable mutation rate at genetic algorithms: introduction of chromosome fitness in connection with a multi-chromosome representation, International Journal of Computer Applications 72.17, 2013

[324] J. He, and G. Lin, Average convergence rate of evolutionary algorithms, IEEE Transactions on Evolutionary Computation 20.2, 2016: p.316-321