Cloud Pricing Models: Taxonomy, Survey and Interdisciplinary Challenges

Caesar Wu, Rajkumar Buyya and Kotagiri Ramamohanarao
Cloud Computing and Distributed Systems (CLOUDS) Lab,
School of Computing and Information Systems,
The University of Melbourne, Victoria 3010, Australia

Abstract

Whether you are a cloud services provider (CSP) or cloud business customers, one of the critical issues is how to understand the cloud pricing strategies and models for either achieving competitive advantages or managing cloud resources effectively to deliver cloud business applications. However, the number of currently available cloud prices and pricing models are overwhelming. It is an intricate issue due to crossing several domains of knowledge, such as cloud technologies, microeconomics, operations research, and even the philosophy of science. It is very challenging even for many cloud experts to fully comprehend the pricing mechanism systematically. Some earlier studies categorized the cloud pricing models via a random walk manner. Others proposed pricing model category in an isolated and unsystematic way. These approaches inevitably lead to much confusion for many cloud decision makers when they need to make the critical decision for business investment. The goal of this paper is to provide a systematic view of cloud pricing in a multidisciplinary approach. This paper presents a comprehensive taxonomy of cloud pricing models that are underpinned by value theory. This systematic approach of taxonomy is driven by three fundamental pricing strategies that are built on the categories of a 3 by 3 matrix. These categories can map out the total of 60 pricing models. Many CSPs have already adopted some of these models. Others have been widespread across many service industries. We provide the descriptions of these model categories and highlight both their advantages and disadvantages. In addition, we put forward an extensive survey of some typical cloud pricing models that were studied by many researchers during the last decade. Based on our survey, we identify four trends of cloud price modeling and the overall direction of cloud pricing, which is moving from intrinsic values per physical box to extrinsic values per serverless sandbox. We conclude that a hyper-converged cloud architecture supported by cloud orchestration, virtual machine (VM), and serverless sandbox will drive the trends of cloud price modeling. We outline four challenges of cloud pricing models and propose some possible ideas on how to deal with these challenges in years to come.

CCS Concepts

CCS → General and reference → Document types → Surveys and overviews

KEYWORDS

Cloud services provider (CSP), Cloud Price Model, Value-based pricing, Market-based pricing, Cost-based pricing

1. INTRODUCTION

Today, the term “Cloud Computing” is believed to be one of the most popular and sometimes an overdosed buzzword that is across many industries. However, the speed of cloud migration from on-premises to off-premises (either private or public cloud) has been very slow, especially for many enterprises or business applications, according to both International Data Corporation or IDC’s and Gartner’s reports [1][2]. There are many issues behind this phenomenon. One of the critical issues is the cost of cloud computing. Many cloud business consumers do not fully understand the cloud pricing [1] in term of the cost. Many Cloud Service Providers (CSP) or cloud advocators argue that at least 68% of the price for “off-premises” is explicit and 32% is unknown whereas “On-premises” only 9% of the cost is explicit while 91% is unpredictable [3]. However, the term “explicit” does not mean cheaper computing. As J. Weinman [4] emphasized that “Cloud computing is not cheap computing...whether the cost-benefit-risk equation favors the public cloud is ultimately unique to each customer. Further confusing things is the challenge of separating arbitrary costs from inherent ones.”

B. Martens et al. [5] echoed this view, and he argued that many cloud cost conclusions lack a systematic approach in evaluating the real costs behind clouds. Many favored claims of the cloud are dependent on the ad-hoc and arbitrary processing of price modeling without the awareness of many indirect and hidden factors. The problem of these claims is to conceive a cloud pricing model as a single discipline issue in an isolated way.

In contrast, Buyya et al [6][7] indicated that the cloud computing and its pricing is an interdisciplinary research issue, which should be studied across cloud technologies (It is to study hypervisors, virtual machine and serverless container, and hyper-converged technologies), price theory (It is regarding of adopting different pricing strategies for various cloud services.), microeconomics (“It is the study of choice under conditions of scarcity” [103]), operations research (It is to use a scientific approach to manage or operate a firm that offers cloud services), and even philosophy of science (It focuses on the foundations, methods, and implication of science). According to [8][9], no single discipline can provide a satisfying solution for cloud pricing because it requires different domains of knowledge from multiple disciplines.

---

[1] Cloud business customers have their own business or applications, 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 business perspective, cloud price that is offered by CSPs is equivalent to their cost.
However, some previous papers attempted to investigate the issue of cloud pricing model in isolation. It leads to an ambiguous classification and misconception of cloud pricing in terms of cloud resource allocation. A typical example is a classification of 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 because of the nature of fluctuation of the price [97][98]. On the other hand, it can also be regarded as market-based or auction-based or resource-based cost-based pricing because of its appearance [99][100]. Consequently, it is inevitable to confuse many decision-makers when they are facing an overwhelming number of cloud prices but have to make the critical business decision to allocate cloud resource for their business applications.

Therefore, this taxonomy and survey will take an interdisciplinary approach to investigate both the pros and cons of different cloud pricing strategies and models regarding a fundamental question of value (or axiology). From the base of the value theory, we investigate some models to reflect the principle of economic value, which is to measure the subjective experiences of cloud services and other models to emphasize the cloud infrastructure cost, and still others to focus on a relationship of supply and demand or the cloud market.

Although we draw multiple disciplines (See Fig.1) for cloud pricing, our investigation will mainly depend on four domains of knowledge: cloud computing (technologies); microeconomics (theoretical tool); operation research (means) and philosophy of science (principle). Briefly, cloud pricing is our goal. Value theory is our principle of logic reasoning, decision theory is our method and microeconomics is our theoretical tool. Cloud technologies underpin the framework of cloud pricing.

![Fig.1: Big Picture of Multiple Disciplines of Cloud Pricing Models](image-url)

Based on Fig. 1, we show that the value proposition (good and bad, right and wrong) is the primary reason for cloud customers to adopt various cloud technologies. The economic sense of subjective value is measured by the quantity or preference of utility, which is an abstract measurement for one’s satisfaction for acquiring the cloud resources. The utility of cloud services determines the price of cloud services that cloud customers are willing to pay. Theoretically speaking, the concept of value theory or axiology is a branch of philosophy. It concerns: “…which things are good or bad, how good or bad they are, and, most fundamentally, what it is for a thing to be good or bad. … What kinds of things are or can be valuable? How can values be compared and measured? How does value theory bear on practical issues?…” [101]

From the axiology [101], we can derive three strategies of cloud pricing through both subjective (values) and objective views (fact). Value-based pricing is demand-driven, and cost-based pricing is supply-driven. Moreover, market-based pricing is the result of the interaction of both supply and demand. Based on these basic strategies, we can articulate a hierarchical framework (as shown in Fig. 2). Each layer of the framework is driven by different goals. At the top layer, the pricing is driven by the principle of value. The next layer down is derived from three pricing strategies, which are to pursue a long-term goal of business applications. The layer further down is drawn from pricing tactical designs, which is oriented by short-term objects. The aim of tactical pricing is how to translate from strategic principles to tactical activities. Finally, the bottom layer of cloud pricing consists of 60 individual models, which is detail oriented. To understand the relationship.

---

2 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.

3 Resource-based pricing is that a firm charges its customers according to resource consumption. Often, the resource will include both materials and time. Many IT projects use this model because the scope of IT project cannot be specified clearly. Many IT professionals or companies quote their resource-based price as IT labor per hour if materials are excluded.

4 Value-based pricing is one of the pricing strategies to offer customers with better values for the products or services. It is based on how much the customers (not a supplier) perceive or are willing to pay. One of the typical examples in practice is the brand pricing strategy, which consumers are willing to pay more for a good brand of a product.

5 Strategy is how does a decision maker deal with or solve the given business problem for a long term or overall goal. There are sophisticated and unsophisticated strategies.

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.
of three pricing strategies clearly, we can leverage the Lerner index formula for profit maximization [10] [11]. (Note, in order to simplify the theoretical discussion of pricing strategies, we assume the firm’s market power is a monopoly). Mathematically, we can derive the following equation:

\[ p = mc + \frac{Q}{\partial Q/\partial p} \]

where “\( p \)” reflects price based on demand and supply and “\( mc \)” represents “marginal cost,” and the last term describes “markup” price. Briefly, this equation reflects the basic principle of economics, which the total revenue equals the total cost plus profits [12]. However, the cost-based pricing does not mean that we can exclude the markup component. Subsequently, the value-based pricing cannot ignore the baseline costs. The equation only illustrates the orientation of the pricing strategies. The implication of this equation represents a relationship of various pricing strategies that were developed by Nagle et al. [13], from later 1980s to present. Nagle’s pricing strategies can be considered as an extension of Monroe’s work [14], which was driven from cost computations, break-even analysis, and target rate of return pricing in the 1970s. If we consider Monroe’s work focused on the relationship between price and quality, then Nagle’s pricing strategies is a shift from Monroe’s cost-based to value-based pricing.

![Fig.2: Orientation of Cloud Pricing Strategies](image)

According the classification of pricing strategies, we can find that currently, many CSPs have adopted various pricing models of either cost-based or market-based pricing strategies because these strategies aim to some cloud service facts or characteristics summarized by NIST (on-demand, broad network access, resource pooling, rapid elasticity, and measured services) [15]. However, these two strategies are insufficient to achieve the maximum profit for CSPs because even if we know all the facts of cloud pricing (such as cloud service cost, markup ratio, market share and target rate of return, etc.) objectively, a pricing decision cannot still be made. The reason is that a CSP does not know how to handle these facts, which fact is more important than the others, why and when it is much more important than others. These questions are the questions of the value. As Karl Popper nicely stated:

“Decisions can never be derived from facts (or statements of facts), although they pertain to facts...It is impossible to derive a sentence stating a norm or a decision from a sentence stating a fact; this is only another way of saying that it is impossible to derive norms or decisions from facts.”[16]

Popper’s statement underpins the principle of our pricing models classification. It means that it would be impossible to derive from the cloud computing facts to the cloud customers’ values. If anyone intends to derive from the fact to the value alone, it becomes a naturalistic fallacy in G.E Moore’s term [17]. Therefore, to investigate and classify cloud pricing models scientifically, we adopt the defined hierarchical framework as shown in Fig.2. As a result, this study made the following contributions for the field of cloud pricing.

- We categorize 60 pricing models into three pricing strategies and nine pricing groups. Many models have not been considered by cloud computing industry yet, but they have been widely adopted by other service industries, such as airline, travel, hotel, recreation, healthcare, telcos, and even retail sectors. The purpose of exploring these potential models is to help many CSPs to compete with their market leaders not only just on price but also on pricing with their business application strengths.
- We survey most of the popular pricing models in a great depth regarding their contributions and gaps plus their business application. Moreover, the paper also unveils the major cloud market leaders often leverage their business strength to model their cloud pricing, which AWS is online retail-oriented pricing. Azure is software-oriented pricing, and Google Cloud Platform (GCP) is search engine optimization (SEO) oriented pricing.
- We identify four research challenges of cloud pricing: 1) How to shift from cost-based to value-based pricing orientation from cloud customer’s perspective, 2) How to move from statefulness to stateless resource pricing and 3) How to shift from mutable to immutable software pricing. 4) how is the cloud pricing modeling developed along with the cloud infrastructure evolution and new technology eruption.
- We also provide some preliminary ideas on how to deal with these challenges in principle.

The rest of the paper is organized as follows: **Section 2** provides the background or a short history of cloud pricing, cloud service launch times, and underlying cloud technologies. In addition, we outline some critical terminologies and their relationship across multiple disciplines based on the principle of economic value. **Section 3** establishes the taxonomy of cloud pricing based on axiology. **Section 4** provides the survey of selected papers that were published from 2009 to 2016. Finally, we compare each price model with other models for its methodology and theory. The conclusion provides three development trends of cloud pricing strategies, which is from cost-based to
value-based, from statefulness to the stateless resource, and from mutable to immutable software pricing. Based on these trends, we highlight four challenges and possible solution.

2. Background

2.1 Cloud Pricing Model Timeline

The first cloud pricing model can approximately be traced back to Salesforce.com’s Russian doll model \[8\] \[102\], which are similar to optimal feature pricing (one of the retail-based pricing models). We can also consider it per-user-based pricing for Software as a Service (SaaS). It is dependent on the value proposition. Salesforce.com’s pricing model is a contrast to Siebel’s outright purchasing or perpetual licensing model. Back in 2000, the average price of CRM software package or system cost around $10,000 per user and $5,000 for the ongoing costs, such as patch, regularly version upgrade, bugs fixing, maintenance, backup, and help desk support. As a result, it was beyond many small and medium enterprises (SME) reach because of their highly volatile business nature. They could not afford to allocate a significant amount IT budget upfront. This issue leads to an opportunity for Marc Benioff to offer SaaS via per-user-based pricing, which Marc stated:

“….to make software easier to purchase, simpler to use and more democratic without the complexities of installation, maintenance and constant upgrades rather than selling multimillion-dollar CD-ROM software packages that took six to eighteen months for companies to install and required hefty investments in hardware and networking.” \[18\]

The cloud technology that underpins the model is so called the software multi-tenancy. The idea of multitenancy is an analog drawing from an apartment building where the tenants can share the cost, such as public facility, building insurance, body-corporation, security, etc. but still have their private space. Similarly, Microsoft Hotmail or Google’s Gmail also offers the email service, which every user (or tenant) can enjoy the email service via any web browser without any stress of installation and configuration of the mail software by themselves. The Fig. 3 summarizes a timeline of different pricing models that were adopted by some leading CSPs along with cloud technologies development.

Following the similar principle, AWS, as IaaS provider adopted the “on-demand” pricing model for its Simple Storage Services (S3) launched in Mar 2006 and Elastic Compute Cloud (EC2) released in Aug 2006. The enabling technology for AWS is Xen hypervisor, which Citrix Systems released the initial version in Oct 2003. Later in 2009, AWS launched an auction-based pricing model or spot instance. Google App Engine or Platform as a Service (PaaS) began to offer a cloud service platform for its customers to host their web applications within the current Google data center in 2008. Its price model is very similar to AWS, which is calculated by cloud instance /per minute or Pay as you Go (PAYG). The underlying software infrastructure for Google Cloud Platform (GCP) to support its PaaS is Kernel-based Virtual Machine (KVM) hypervisor that was initially released by Qumranet in 2006, but Red Hat now owns it. Red Hat KVM is open source software. The other leading CSP player is Microsoft Azure. It began its cloud service business in Jan 2010. The price model is almost identical with AWS. Although Azure entered the cloud service market relatively later, it has quickly captured cloud market share according to Gartner’s Magic \[2\]. Azure’s cloud service technology is MS Hyper-V. Microsoft launched Hyper-V in 2008. As we should see, the top three leading CSPs use three different hypervisors, but a significant proportion of public CSPs adopt Citrix Xen, such as Softlayer, Rackspace, GoGrid, Oracle VM for x86, Aiyun (Both Xen and KVM) and Virtustream’s µVM (or Microvisor) adopts Xen. Linode moved its VMs from Xen to KVM in Jun 2015 because it believes that KVM is 28% faster than Xen. However, the much common hypervisor for many private clouds is VMware, which is the first commercial hypervisor that was launched in 1999.

\[8\] 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.
Some public CSPs, such as CenturyLink and Interoute also adopt VMware as its hypervisors to support their cloud business because VMware provides a comprehensive toolset that allows customers to manage their private cloud service efficiently. However, if host applications are moved to a public cloud, it would become too heavy and cumbersome. One of the significant disadvantages for VMware is overpriced. Another interesting observation is that most public CSPs adopt Xen hypervisors and their pricing model is on-demand based or Pay As You Go (PAYG). The minimum billing unit is per hour base. On the other hand, if CSPs use KVM hypervisor, the billing unit is reduced to per minute. Some CSPs that adopt VMware often require customers to have a long-term commitment in term of a service contract. (Refer to Table 1

<table>
<thead>
<tr>
<th>Name of CSP</th>
<th>Type of Cloud Models</th>
<th>Initial offering time</th>
<th>Minimum billing Unit/Cycle</th>
<th>Type of Cloud Service</th>
<th>Supported Technologies</th>
<th>Hypervisor Launched Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salesforce.com</td>
<td>Per User Based</td>
<td>1999</td>
<td>Monthly</td>
<td>SaaS</td>
<td>Multitenancy</td>
<td>-</td>
</tr>
<tr>
<td>AWS</td>
<td>On-Demand</td>
<td>2006</td>
<td>Hourly</td>
<td>IaaS</td>
<td>Xen</td>
<td>2003</td>
</tr>
<tr>
<td>AWS</td>
<td>Auction Based</td>
<td>2009</td>
<td>Hourly</td>
<td>IaaS</td>
<td>Xen</td>
<td>2003</td>
</tr>
<tr>
<td>Google App Engine</td>
<td>On-Demand</td>
<td>2008</td>
<td>Minute</td>
<td>PaaS</td>
<td>KVM</td>
<td>2006</td>
</tr>
<tr>
<td>Google Cloud Platform</td>
<td>On-Demand</td>
<td>2010-2014</td>
<td>Minute</td>
<td>IaaS</td>
<td>KVM</td>
<td>2010</td>
</tr>
<tr>
<td>Azure</td>
<td>On-Demand</td>
<td>2010</td>
<td>Hourly</td>
<td>XaaS</td>
<td>Hyper-V</td>
<td>2008</td>
</tr>
<tr>
<td>Softlayer (IBM)</td>
<td>On-Demand</td>
<td>2006</td>
<td>Hourly</td>
<td>IaaS</td>
<td>Xen</td>
<td>2003</td>
</tr>
<tr>
<td>Rackspace</td>
<td>On-Demand</td>
<td>2008</td>
<td>Hourly</td>
<td>IaaS</td>
<td>Xen</td>
<td>2003</td>
</tr>
<tr>
<td>GoGrid</td>
<td>On-Demand</td>
<td>2006</td>
<td>Hourly</td>
<td>IaaS</td>
<td>Xen</td>
<td>2003</td>
</tr>
<tr>
<td>Aliyun</td>
<td>On-Demand</td>
<td>2012</td>
<td>Hourly</td>
<td>IaaS/PaaS</td>
<td>Xen and KVM</td>
<td>2003, 2006</td>
</tr>
<tr>
<td>Virtustream (Dell)</td>
<td>Per User Based</td>
<td>2012</td>
<td>Monthly</td>
<td>IaaS</td>
<td>Microvisor (Xen)</td>
<td>2012</td>
</tr>
<tr>
<td>Joynet</td>
<td>On-Demand</td>
<td>2013</td>
<td>Minute</td>
<td>IaaS/PaaS</td>
<td>SmartOS</td>
<td>2011</td>
</tr>
<tr>
<td>Linode</td>
<td>On-Demand</td>
<td>2008</td>
<td>Monthly</td>
<td>IaaS</td>
<td>From Xen to KVM</td>
<td>2003, 2006</td>
</tr>
<tr>
<td>CenturyLink</td>
<td>Reserved Based</td>
<td>2011</td>
<td>Monthly/Yearly</td>
<td>PaaS</td>
<td>VMware</td>
<td>1999</td>
</tr>
<tr>
<td>Interoute</td>
<td>Reserved Based</td>
<td>2012</td>
<td>Monthly/Yearly</td>
<td>IaaS</td>
<td>VMware</td>
<td>1999</td>
</tr>
<tr>
<td>Oracle VM for x86</td>
<td>Reserved Based</td>
<td>2012</td>
<td>Yearly</td>
<td>IaaS/DeaaS</td>
<td>Xen</td>
<td>2003</td>
</tr>
</tbody>
</table>

The key threads of cloud price modeling are on-demand and reserved pricing models. Spot instance is more likely as a sideshow or marketing campaign tool to prompt AWS cloud services because spot instance cannot support the mission-critical application. And yet, many B2B applications require the mission-critical infrastructure. The basic logic behind the reserved pricing model is for cloud resource certainty, which is basically cost-account driven. On-demand focuses on the flexibility of cloud resource allocation, which is the value-based pricing.

### 2.2 Crucial Terminologies and Logic Relationship of Cloud Pricing

Due to the interdisciplinary nature of the price modeling, it is essential to clarify some fundamental terminologies that will be adopted in the later sections, namely value-based, market-based, cost-based pricing strategies, revenue or yield management, an industry organization, marginal cost, and markup. These terms are the foundation of our taxonomy and survey of cloud pricing models.

#### 2.2.1 Value-Based Pricing

The word “value” means how much worth to an agent for an object. It is measured by a unit of utility. Fundamentally, value concerns things are good or bad in a successful and efficient sense [19] [20]. 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 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 or additional value due to new cloud characteristics), and 3) “Good to be” (e.g., the strategy is to promote cloud customers to migrate more workloads to off-premises, which is to identify future value). We can briefly illustrate the relationship between three types of good values in the following Fig.4 and have a further explanation in section 3.1.

In comparison with other pricing strategies, value-based pricing is much subjective. It might not be necessary to reflect in a market price and service costs fully. A typical example is perception-value, which is based on the customers’ perceptions of what is expected in comparing 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 the pricing to be paid [21]. According to Sheth et al. [22], customers perceived values has 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 [23], such as emotional and epistemic. However, it is quite challenging to be constructed because of subjectiveness, especially for business to business (B2B) type of service because “perceived values” are the primarily measured satisfaction of the individual customer. With the B2B type of service [24], it is hard to detect the real users' satisfaction directly. Instead, the perceived values are contributed by an indirect person, such as a manager’s or decision maker’s perception.
Overall, the value-based pricing strategy consists of three groups that have at least 14 different pricing models discussed in section 3.1. Customer demands are often the primary factor that impacts these models. The criteria to differentiate these models are the representation of current, new and future values from a customer’s perspective. Strategically, these models emphasize on customer’s experience, satisfaction, expectation, and perceiving measurements. Furthermore, we can extend the value-based criteria to both market-based and cost-based pricing. By doing so, we can form a 3 by 3 matrix as the classification criteria of cloud pricing models. (Refer to Table 2)

**Fig.4 Three Types of Good Values In term of Value-Based Pricing**

<table>
<thead>
<tr>
<th>Axiology/Pricing Strategies</th>
<th>Value-Based Pricing</th>
<th>Market-Based Pricing</th>
<th>Cost-Based Pricing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good to Have</td>
<td>Good to Have + Value-Based</td>
<td>Good to Have + Market-Based</td>
<td>Good to Have + Cost-Based</td>
</tr>
<tr>
<td>Good to Do</td>
<td>Good to Do + Value-Based</td>
<td>Good to Do + Market-Based</td>
<td>Good to Do + Cost-Based</td>
</tr>
<tr>
<td>Good to Be</td>
<td>Good to Be + Value-Based</td>
<td>Good to Be + Market-Based</td>
<td>Good to Be + Cost-Based</td>
</tr>
</tbody>
</table>

### 2.2.2 Market-Based Pricing

“Market-based pricing” does not only weight single CSP but also compare with other CSPs in the current market for the offered services and associated prices in term of the supply curve. Therefore, it is also called competitive pricing. The market is the intersection place where all customers meet all CSPs [25]. The market environment will fertilize 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 3.4) “Freemium” is to give away a product with basic functionality or features for free to gain the market share.

“in an environment of very low or no marginal distribution and production costs that provides the potential for massive scale, with advanced functionality, premium access, and other product-specific benefits available for a fee” [26]

The main purpose of this pricing 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. Practically, CSP may adopt different pricing models to implement its 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. [27] highlighted four different measurements, which they are features, quantity, quality, and period. These models can attract many high-end customers and get many valuable feedbacks from a large number of audiences for a service provider to improve the new products or services. One of the main risks is that these free trial customers might never be converted to premium customers due to intensified competition in the marketplace. The key to this model is how to draw a line between free and premium. Overall, market-based pricing strategy includes three types of pricing categories, namely “Free and later pay,” “Retail-based pricing,” and “Auction and Online based pricing.” However, the number of pricing models is up to 35. Retail-based pricing contributes the majority of the models (26) in this category which we discuss in section 3.5. The criteria to classify these market-based pricing models can be summarized as two guidelines: one is to examine what is the orientation of these models in term of values, and two is to see whether can be validated by the cloud market (external validation).

### 2.2.3 Cost-Based Pricing

Most of the enterprises and government agents with on-promises cloud resources adopt cost-based pricing because it is much easier to be understood 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 can also be considered as fact-based pricing. Although many pricing experts emphasize value-based pricing [28] [29] [30], the cost-based pricing can help decision makers to set a baseline to charge customers for the minimum price so that they can at least cover all the production expenditures. Another benefit of cost-based pricing is it can articulate the unit cost and provides a measurement for a benchmark comparison. It becomes one of the managerial tools for the decision maker to drive the performance or price competitiveness. Lastly, the components of cost are the essential element of Cost and Benefit Analysis (CBA) so that a decision can be made much realistic [31]. The cost-based pricing strategy comes with different categories. Based on the value of cost account, we can have three categories and 11 pricing models. The expenditure based pricing focuses on the real cost. Resource-based pricing is often applied to business
to business (B2B) scenarios. On the other hand, utility-based pricing is typically designed for business to consumer (B2C). From a baseline perspective, the cost-based pricing focuses on internal validation. In practice [32], value-based pricing often remains as a minority, and market-based pricing is the dominated pricing models followed by cost-based pricing strategy as shown in Fig.5. It should not be surprised because many value-based assessments are often unreliable due to the subjectiveness of customers’ perception. Based on Emil Kauder’s historical observation [33], British economists lean to cost-based pricing theory while Frenchmen and Italian scholars prefer the value-based pricing theory before the 19 century.

Contemporary scholars, Nagle et al. [13] proposed the solution of value metrics for the value-based pricing. They articulated six activites or value cascade to implement value-based pricing, which they are value creation, value communication, price structure, pricing policy, price setting, and price competition. Nagle’s practical technique should help many cloud practitioners to implement the value-based strategy.

![Summary of Different Cloud Pricing Strategies in Practice](image)

**Fig.5 Adoption of Pricing Strategies in Practice [34]**

2.2.4 Economic and Managerial Terms

Price theory, market theory, and industrial organization are all a sub-discipline of economics. Operational Research and Decision-Making theory is the sub-discipline of managerial science. There are also some terms emerged from inter-subdisciplines, such as strategic managerial accounting, managerial economics or even quantitative analysis for competition. Many terminologies are adopted by cloud pricing models. They are less familiar with computer scientists. So, it is essential to clarify these terms up front.

2.2.4.1 Yield and Revenue Management

The term “yield management” means to maximize the sales revenue yield or profit by an optimized approach for both selling price and sales volume or quantity. Generally, yield management has a narrow focus [35]. Practically, yield management is a part of revenue management. For example, if a CSP offers IaaS, the CSP should find a trade-off balance between prices of per virtual machine and the volume of VM. On the hand, “revenue management” has a broad scope or look at the big picture. It considers an optimized approach to generate or maximize multi-yields or different revenue streams for additional services’ sales. When we join these two words together, it means to maximize sales revenue from various service streams. However, we should be careful about the term of revenue maximization because it does not necessarily mean profit maximization. A pricing strategy is possible to generate the maximum revenue, but it can still lose money or profit if the average total cost is higher than the average sales price.

2.2.4.2 Perishable Services

“Perishable services” refers to the services that have to be consumed or sold within a limited period. The common perishable services are airline or concert or sport event tickets, broadcast advertising time slots, hotel rooms, and healthcare. The essential characteristics of perishable services are they cannot be kept or accumulated in stock. If we combine both terms of perishable assets and revenue management, the interdisciplinary term is emerged, which is called perishable-asset revenue management (PARM). According to Weatherford et al. [36], the PARM has at least three characteristics 1) the products and services that derive from these assets could be available at one date, and it could be either unavailable or age after. 2) The capacity of these assets is fixed. An incremental capacity will require a higher cost and time buffer. The cloud data center capacity is one of the typical examples. In comparison, the variable cost is relatively lower than a fixed cost. For example, the variable cost for an additional unit of cloud instance is only $0.1 when the initial fixed cost to provide the cloud service is often $5. So, it would be preferable to have a substantial discount on cloud instance price rather than let the cloud capacity running in idle. 3) Perishable services have the possibility of segmenting price-sensitive customers. It is because a particular service has many service features or characteristics can be used to segment the market. Another common practice for market segmentation is the time of purchase (or Service on Demand). For example, the price of an airline ticket is much higher for business customers because they often purchase the tickets at the last minute. In comparison with price-sensitive customers, they usually make a ticket reservation far earlier but without any flexibility. AWS adopts a similar pricing strategy, such as dynamic pricing [37] for its yield and revenue management [37] or PARM.

2.2.4.3 Cost, Price, Marginal Cost, Markup price, Profit and Net Present Value (NPV)

The concept of cost is similar to a notion of “price.” From a customer’s viewpoint, it is a payment or expenditure for services that can be obtained or delivered. From a CSP perspective, it is the amount of payment for time and material to deliver cloud service for a customer before the profit margin. There are different types of cost, namely, fixed cost, variable cost, average total cost, sunk cost, and marginal cost.
The fixed cost does not vary when the volume of producing services is increasing. Variable cost, on the other hand, is rising when the volume is up. The average total cost is the sum of fixed and variable costs divided by the total number of the product. Marginal cost is the changing costs divided by changing quantity to be produced. It is an incremental cost for each unit to be produced. The price occurs during the process of transaction between seller and buyer. Sometimes, it is also written as a sum of marginal cost and markup price. It is also called an absolute or sales price or market value. It is to measure how much worth for a purchased service in comparison with other services at the marketplace. On the other hand, the markup price is the ratio between sales price and cost. It is a relative price. It can be presented as “markup price = (sales price/cost) -1”. Profit is equal to sales price minus a cost. From a broad perspective, it is the total revenue minus the total expenditure. Tactically, the pricing model intends to achieve the profit maximization. Strategically, the pricing model is the context of both yield and revenue management under some conditional constraints, such as resource capacity, budget, and time. Net Present Value (NPV) is one of the financial concepts, which is to measure the present value of assets from a future perspective or after a particular period T (for example, five years) due to the capital cost (or interest or discount rate, “r”). Refer to the following equation, where NCF_t is the Net Cash Flow at each year (or t = 1, ..., 5) [39]

\[
NPV = \sum_{t=0}^{T} \frac{NCF_t}{(1 + r)^t}
\]

3. **Taxonomy of Pricing Models**

The pricing strategies consist of 60 different pricing models from four aspects namely value, fact, supply, and demand. From these four aspects, we can derive a comprehensive taxonomy of cloud price models that include 9 different categories that are formed a 3X3 matrix as shown in both Fig.6. Each category of pricing has between 3 and 6 pricing models except retail-based pricing models. In the following sections (from 3.1 to 3.9), we will first introduce and define each category of pricing models. We then, will explain why some models have been adopted by CSPs during the last decade or so, and other not in the summary section.

![Fig.6: Taxonomy of Overall Cloud Pricing Models](image)

![Fig.7: Taxonomy of Retail-Based Pricing](image)

3.1 **Customer Value-Based Pricing**

This category of value pricing consists of four pricing models, namely perceived-value, psychological, feature, and hedonic based pricing. A customer’s value proposition is the main reason for this category of pricing. If the customers believe the value is “good to have” rather than not, they will be willing to pay (W2P) for it. These four models are constructed by the value metrics of the cloud services in 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 because of W2P. The main challenge is how to define the value metrics by measuring customers’ subjectiveness value for “good to have.” In comparison, both feature and hedonic based pricing models are relatively objective in comparison with the other two models. These pricing models are suitable for an ever-changing environment in term of new cloud features. However, not every feature or characteristic of service would be “good to have” for every customer. Some features might have negative values or burdens to some customers. As a result, the selection of specific service features for pricing model becomes subjective or challenging.

### 3.2 Experience-Based Pricing

This pricing category can be distinguished as the value is “good to do.” It is measured by customers’ performance experiences, such as specified reliability of a cloud service or utilization rate of a limited resource (e.g., cloud infrastructure). The aim of these models is to increase the new values what the customers had already have. To some extent, it can also be considered as performance-based pricing. According to M McNair 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 in the service industry is online advertising payment, which is dependent on the measurement data, such as the number of clicks or purchases [41]. Other applications are telecom services (such as multi-party video conference, mobile apps, satellite connectivity, etc.), which the service prices rely on its specified performance metrics. This pricing model category is dependent on the buyer’s business outcome. The basic idea is to make sure that a seller’s products or service meet the buyer’s business objectives or value. The reward of this model is that both parties’ values are aligned. By doing so, the seller will not undercharge the pricing and buyer 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 vendor. 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 advertising camping? Sometimes, the advertising campaign time may take longer than what was originally 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 “customer value-based pricing category, the value metrics appear to be much tangible.

### 3.3 Service-Based Pricing

Service-based pricing category is to emphasize the value of “Good to be” for cloud customers. In comparison with the above other two value-based pricing categories, the value of “good to be” focuses on the future values. Many CSPs of SaaS adopt service-based pricing models, such as Salesforce.com and MS Azure. It can also be considered as incentive-based pricing because this pricing category is based on 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. Every “good to be” value is incremental. There are six different ways (or pricing models) of value calculation: on-demand, tier based, per user based, per device based, all you can eat, and priority-based pricing. The driving forces of these models are often a Service Level Agreement (SLA) [63] or Operation Level Agreement (OLA) [91]. Overall, the service-based pricing is the direction of future value metrics, which is “on-demand.”

The concept of service-based pricing is often 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-add while resource-based pricing emphasizes the requirement of various inputs.

The disadvantage of per user or per device models is when the number of users or devices is more than the particular threshold level the price models would become unfavorable from a customer perspective. For example, if the number of CRM software users is too large, then it would be good to purchase Siebel license outright at a bargain price rather than pay usages by per user base. The other hurdles are how to quantify tiers or priority. Sometimes, it is not clear. For instance, AWS’s network performance tiers for a cloud instance are specified in “very low,” “low,” “low to moderate,” “moderate,” and “high.” From a customer perspective, the values of these tiers are quite vague because a business metrics often require identifying in Mbits or Gbits, not low or moderate or high.

### 3.4 Free up Front and Pay Later Pricing

Due to intense competition, many firms, especially, e-commerce businesses adopt “Free up front and pay later” pricing model. The idea is to leverage free products or services with minimum features or characteristics 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 ads model, it can stimulate customer demand, and customers can enjoy free products or services. The bad news is the customer may waste a lot of unnecessary time. Moreover, this model requires the particular size of market share from a CSP perspective. If the market size is not large enough to offset the cost of free products, the pricing model is unsustainable. Google’s 22 free online services and MS Office online are typical examples of this kind of pricing. For the freemium one, 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 measurements of freemium, which are quantity, a 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 clicks and memory size/per month).
In term of Razor-and-Blades model, it 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 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 the market share.

### 3.5 Retail-Based Pricing

By its name, the retail-based pricing models are based on a small quantity that consumers buy from fixed locations or retail outlets (such as online portal, milk bar, shopping malls, petrol station, department stores, supermarket, Sunday market, etc.) By and large, the retail provider should 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: 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 Fig.7, which the products nature drives the product mix pricing, the payment option guides the discounts and allowances pricing, the sale strategy orient the promotional pricing, and the customer segment mainly dominates the discriminatory pricing (Refer to Fig. 7).

#### 3.5.1 Product Mix Pricing

This pricing model is to mix or combine with different types of pricing schemes in different ways. In essence, there is no single pricing model to lock in consumers for their resource demand. They can depend on their usage pattern to combine different pricing models. The standard practice for the cloud customer is to combine fixed and on-demand 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 to how 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. From a supplier perspective, it may increase the operational cost. The bottom-line is how to make a rational decision on pricing that can reflect customers’ demands by different segments.

#### 3.5.2 Discounts and Allowances Pricing

Price discounts and allowances are two techniques for a firm to response fluctuation conditions due to a dynamic market. 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 is one of the perishable assets. AWS had 40% 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 small business. 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 primary driver behind this subcategory of pricing models is payment or the idea of 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 or sales volumes, especially for perishable assets, like cloud assets. The main disadvantage of these models may reduce profit margin and do not have a brand identity.

#### 3.5.3 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 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” [43]

The purpose of promotional pricing is to boost sales, reduce stock, stimulate demand and maximize business opportunities. The obvious reward is to increase sales and minimize stock level [44]. The unfavorable point is 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 because a manufacturer would like to reduce the stock inventory for a particular model of the laptop due to Moore’s law effect.

#### 3.5.4 Discriminatory Pricing

Discriminatory pricing means that the pricing scheme is charging different prices for 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 [11], it is the 1st degree of pricing discriminatory, which is usually dependent on one-to-one negotiation, such as property sale (in private
It often requires a lot of efforts and challenging to capture the customer’s maximum value. It is not common to apply to commodity products. 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 citizen, students, it is the 3rd degree of price discriminatory. For instance, Microsoft charges student license for MS office package. If we combine different types of price discriminatory, we should have various price schemes 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 the customer segmentation, which is to design different pricing schemes for various groups of customers. Amazon segments its customers by mixing of operational revenue streams [45] as shown in Fig.8. This pricing subcategory does not only allow a provider to boost its sales but also to maintain the profit margin. The flip side of these pricing models would increase sales cost and increase a stock inventory level, which will ultimately increase the investment risks. In this category of pricing models, the criteria of model classification are two measurements: market-based strategy or market segmentation and value principle of “Good to do” to create new values for CSPs.

![Graph showing AWS vs Non-AWS Segments Revenue YoY Growth](image1)

### 3.6 Auction and Online-Based Pricing

#### 3.6.1 Auction Pricing

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 of the auction-based pricing is that the market price of some products, such as artworks, antique 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 billion-dollar mobile spectrum. Many commodity products, property, and financial bonds are included. AWS also places its EC2 and S3 under the auction bidding rules.

There are some pros in term auction-based pricing: The speed of the auction is relatively fast. It is clear that 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 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 real auction pricing models can be categorized into four models: Spot and forward pricing, English Auction, Dutch Auctions, and Sealed-bid Auction. This paper only focuses on a few auction schemes that are closely associated with the cloud computing market. For example, AWS has adopted a modified spot and forward pricing since 2009. The term “spot” literally 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, only AWS provided the spot instance for public cloud customers.

#### 3.6.2 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 offer both online and offline purchasing prices for customers, but the offline price may be higher than the online one. For example, Officeworks provides both online and offline prices, but the offline price is sometimes higher than online. Amazon only offers online price for its retail goods and services.

The upside of online pricing is it can instantly reach a vast number of customers. 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 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 issue, lack of or no significant discount, frauds in online pricing and the extra cost of goods delivered.

### 3.7 Expenditure-Based Pricing

Expenditure-based pricing means every price model is derive 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 subcategory is all pricing models are proportion to the particular percentage of the total cost. These models are often preferable schemes for many enterprise firms and government agents.

The benefits of these types of models are that a CSP knows what it wants. They are very concise, more straightforward and quick to be constructed. They can guarantee the profit 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 cost base is inaccurate, it would lead to a wrong pricing model. 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 service delivery. As a result, larger firms or enterprises may lose a lot of business opportunities. The category of cost-based pricing is mainly to consolidate the current value from the CSP perspective.

### 3.8 Resources-Based Pricing

Instead of pricing on cost account, resource-based pricing focuses on a consumption base. The cost-based and resource-based pricing have some common properties that are associated with the expenditure components. That is why we classify resource-based pricing as a subcategory of cost-based pricing. However, not all resource consumption costs money. Some natural resources are free. For example, the natural solar or wind resources do not cost any money. The resource-based pricing emphasizes on scalability, which means that the expenditure inputs are uncertain in order to achieve the desired output. In contrast, the cost-based pricing can often establish a tied cost relationship between inputs and outputs. Many cloud services require resource-based pricing, such as business intelligence analysis at the backend. The other key difference between cost-based and resource-based pricing is that the later one includes both tangible (e.g., physical infrastructure) and intangible resources (e.g., brand name, know-how, standard, procedures, intellectual property, and CSP’s reputation).

If cost-based pricing is mainly applied to manufacture industry, then resource-based pricing can be considered as a pricing scheme for the services industry. Traditionally, there are many service industries adopted resources-based pricing models, such as e-commerce, airline, travel and leisure, recreation and entertainment, healthcare, and education. Resource-based pricing is widely adopted by the IT industry, especially for IT outsourcing purpose.

Resource-based pricing aims to offer a better method that allows customers to consume and deploy the scalable resources both efficiently and effectively. From a cloud computing perspective, resource-based pricing reflects two of five characteristics of cloud computing, namely on-demand and resource pool. In comparison with other market-oriented pricing models, which intend to balance supply and demands, resource-based pricing could be used to artificially overcharge services to eliminate the over-consumption of the scarce resources. The apparent obstacle is to reduce sales quantity and revenue.

In summary, this category emphasizes resource scarcity. There are four types of resource-based pricing, namely, Transaction based, FTE-based, Licensing-Based, Time-Material Based pricing. We can differentiate this category with other pricing models with two criteria: cost-based pricing and value of “Good to do.”

### 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 served different connotations. From a computer software perspective, it means that the software can perform different specified tasks or functions. For example, utility software (iOS or Windows) can be utilized to perform the tasks of monitor, mouse, printer and disk driver. From a public service perspective, it means an incumbent service provider can provide utility services, such as telecom, electricity, gas, water, public transportation, which these services are essential to the modern society. The economic term of utility is that the person receives the total satisfaction or pleasure for consumer goods or services. The philosophical or ethical meaning of utility was defined by English philosopher, Jeremy Bentham [48] in which an action is right because it can produce the maximum amount of pleasure for the maximum number of people.

From a cloud pricing perspective, the connotation of utility pricing is similar to the metered price. 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, the upfront capital expenditure. The flipside of this model is that each has to rely on a CSP for the cloud services.

According to a combination of business requirements, usage time, resource commitment, customer segments, and payment types or different consumption patterns, utility-based pricing can have different pricing models, namely, Peak and Off-Peak and fixed cost based pricing. The criteria for this type of pricing are the combination of cost-based pricing and value of “Good to be.”

### 3.10 Summary of Pricing Models Classification
Cloud pricing strategies, tactics, and models are largely dependent on new features of cloud services, cloud technologies, and cloud orchestration. Based on both Fig. 3 and Fig. 6, we can see that the service-based pricing category, especially on-demand, per use-based and tier-based pricing models became the dominated pricing models for many CSPs, such as Salesforce.com to delivery SaaS in 1999 and AWS to delivery IaaS in 2006. The driven force behind these pricing models is that they can capture one of the fundamental characteristics of cloud service, which is PAYG. In contrast, many cost-based pricing models, such as expenditure base, resource-based pricing models that were popular before the cloud era (dedicated hosting) have been gradually faded away.

The trend of cloud pricing model is moving towards a much more flexible direction and becomes much competitive rather than upfront pay pricing models. For example, AWS introduced the new cloud service features, namely Lambda function as a full serverless platform in Nov 2014. The corresponding function-based pricing model is similar to “Free and Later Pay.” However, there are some slight differences, which the “free” part is not a physical component but a quantified service capacity with a specified timeframe. Clearly, the new pricing model is to support the new service (FaaS) that is working with new orchestration, such as AWS Cloud Watch and Google’s open platform - Kurnebets. Following AWS leadership, both Google Cloud Platform (GCP) and Microsoft Azure also launched Functions as a Service (FaaS) platform in early 2016. All three CSPs have almost identical pricing model for serverless computing (See Table 3). (The detail how to calculate the price for FaaS will be illustrated in section 4.4)

### Table 3. FaaS Pricing Model

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

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

Practically, we can have at least 60 different pricing models to apply to cloud computing. These models can be further classified into nine categories. The detail of each pricing model is excluded from this paper due to the limited space. Our analysis result is similar to Hinterhuber’s findings shown in Fig. 5, which the dominated pricing strategy is market-based pricing strategy (35 pricing models). Although there are many proposed value-based pricing strategy, only a few firms adopted it due to its subjectiveness and challenging to be implemented. The most popular pricing strategy for many large enterprises is a cost-based pricing strategy because it is simple, objectiveness, and direct.

We have defined and highlighted many pricing model categories that have been already applied to different industries, especially service industries. Although many of them have not been adopted, a CSP should not eliminate itself to the limited number of pricing models. Weinman [4] indicated CSP should learn from other industries and compete on pricing, not on price alone. The following table 4 provides the summary information of these categories of pricing models at a glance.

### Table 4. Summary of Pricing Models

<table>
<thead>
<tr>
<th>Name of the model category</th>
<th>Qty. of models</th>
<th>Sub-C Qty.</th>
<th>Simple Definition</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Typical example of Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Value Based</td>
<td>4</td>
<td></td>
<td>The pricing models are driven by customers’ value proposition of “good to have”</td>
<td>Maximized the sales profit based customers W2P</td>
<td>Challenging to select the right service features for pricing models</td>
<td>Many services real including real estate industry</td>
</tr>
<tr>
<td>Experience Based</td>
<td>4</td>
<td></td>
<td>The pricing category is driven by customers’ value proposition of “good to be.” It is equivalent to performance-based pricing</td>
<td>It is a win-win model and fair to both parties</td>
<td>Not every service can be specified with a list of performance metrics</td>
<td>On-line advertising camping</td>
</tr>
<tr>
<td>Service Based</td>
<td>6</td>
<td></td>
<td>It is driven by the customer value proposition of “good to do” (select)</td>
<td>The value can be defined objectively</td>
<td>If the quantity grows fast, the cost could be out of control</td>
<td>SaaS delivery</td>
</tr>
<tr>
<td>Free and Later pay</td>
<td>3</td>
<td></td>
<td>It is to leverage free products with minimum features for generating higher profits from premium customers</td>
<td>Increase customer base and market share</td>
<td>Challenging to decide product components between free and premium</td>
<td>E-commerce, pay TV, proprietary software license</td>
</tr>
<tr>
<td>Pricing Model</td>
<td>Type</td>
<td>Sub-RB</td>
<td>Description</td>
<td>Advantage</td>
<td>Disadvantage</td>
<td>Industry</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>------</td>
<td>--------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>--------------------------------------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Retail-Based</td>
<td>26</td>
<td></td>
<td>It is a B2C type of pricing model</td>
<td>The optimizing product set to maximize profit</td>
<td>Too many options</td>
<td>Retail industry or online retail</td>
</tr>
<tr>
<td>Sub-RB: Product Mix</td>
<td>6</td>
<td></td>
<td>Product-oriented pricing models</td>
<td>Boost sales, generate extra revenue</td>
<td>Lead to a bad reputation</td>
<td>Telco services</td>
</tr>
<tr>
<td>Sub-RB: Discounts</td>
<td>6</td>
<td></td>
<td>Payment driven pricing models</td>
<td>Increase cash flow</td>
<td>Reduce the profit margin</td>
<td>Nearly all retail industries</td>
</tr>
<tr>
<td>Sub-RB: Promotional</td>
<td>7</td>
<td></td>
<td>Sales strategy driver pricing models</td>
<td>Increase sales and reduce inventory stock</td>
<td>Reduce the overall profit margin</td>
<td>PC sales</td>
</tr>
<tr>
<td>Sub-RB: Discriminatory</td>
<td>7</td>
<td></td>
<td>Customer segmentation driver pricing models</td>
<td>Increase profit margin</td>
<td>Increase sales cost</td>
<td>Service industries and automobile retail</td>
</tr>
<tr>
<td>Auction</td>
<td>5</td>
<td>4</td>
<td>Price is settled by bidding based rules in public</td>
<td>Price is transparent; Price is quick to be set down</td>
<td>The price is unpredictable</td>
<td>Real estate industry</td>
</tr>
<tr>
<td>Online</td>
<td>1</td>
<td>3</td>
<td>Price is published on a web page</td>
<td>No extra handling cost, Price is transparent</td>
<td>High risk of Security and privacy issues</td>
<td>e-commerce</td>
</tr>
<tr>
<td>Expenditure-based</td>
<td>3</td>
<td>4</td>
<td>Price is decided by a proportion of cost or expenditure of production</td>
<td>Easier to be constructed and understood</td>
<td>Either overshoot or undershoot</td>
<td>Dominated firms often have a market monopoly</td>
</tr>
<tr>
<td>Resource-based</td>
<td>4</td>
<td>7</td>
<td>Price is decided by resources to provide the services</td>
<td>Consumers deploy scarce resource both efficiently and effectively</td>
<td>Providers have no incentive to optimize price</td>
<td>Professional Consulting industries</td>
</tr>
<tr>
<td>Utility-Based</td>
<td>4</td>
<td>5</td>
<td>Price is metered. Usage is monitored. Payment is according to usage or pre-defined plan in a regular term</td>
<td>Each customer can access the service that is unfordable by a single individual</td>
<td>Each individual has to rely on the utility service</td>
<td>Utility industries: gas, electricity, water supply, and sewage, telco</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Although we listed 60 models in the above, the driving forces behind these models are three pricing strategies, which they are value-based, cost-based and market-based pricing. These strategies become our baseline or principles to classify different pricing models. The reason to have so many different types of pricing models is that different products or services require a different approach to address various issues of service, payment, promotion, and discrimination. As we can see, nearly 50% of pricing models are retail-based pricing. Cloud service pricing can learn from other industries.

Throughout the taxonomy of pricing models, we emphasized on value-based pricing strategy for cloud service because the nature characteristics cloud computing is service oriented. However, it does not mean that the cost-based pricing is not important. The cost-based pricing strategy is still important. It often provides a bottom line price for CSPs. In contrast, the value-based pricing will illustrate 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 price 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.

In general, the cost-based pricing only provides the fact of cost accounts, but it does not tell which cost is more important than others and when it is important because all cost accounts are equal. The value-based pricing, on the other hand, can display the value proposition. The relationship between value-based and cost-based pricing strategies are additive. This relation leads to a 3x3 metrics of value measurement for our classification criteria based on a strategic goal.

### 4. Survey of Pricing Models

During the last two decades 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), and utility computing. Based on our classification criteria, the following survey will be organized into three cloud pricing strategies. We selected published work between 2009 and 2017 for detail investigation. One of the compelling reasons for selecting these studies is the majority of the research works presented their finding or price scheme is either in a new mathematical model or a novel algorithm.

According to the context of these papers, we classify papers [55][56] [61][62] as market-based pricing and papers [60] [77][78][80] as cost-based pricing and the papers [84][85][86][91][93][95] as value-based pricing. We highlight the uniqueness of their ideas, new concepts, and the contributions of each paper. Moreover, we show their relationship, whether it is a continuation of previous work or the original work.

#### 4.1 Pricing Models of Pre-Cloud Computing

In later 1999 and early 2000s, Buyya et al. [49], [50] presented a computational economy framework to regulate grid computing resource based on market supply and demand. The basic idea of this structure is 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, namely commodity market, posted price, bargaining, tendering/contract-net, auction, bid-based proportional resource sharing, community/coalition/bartering and monopoly and 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 for a cloud computing paradigm.

When virtualization has become a mature technology in the middle and later 2000s, cloud computing is on the horizon. Based on many years’ research experiences, Buyya et al. [52], [53] argued that the paradigm has 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 (such as a cloud broker) to allow cloud consumers to utilize global cloud infrastructure. The idea of global cloud or multi-cloud service providers was a very cutting edge at that time. It has only become practicable after the serverless container technology has emerged recently.

### 4.2 Market-Based Cloud Pricing

By leveraging Buyya 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 to reach the highest revenue, 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 large-scale simulation dataset provided by Google. The following four equations can illustrate the essence of their solution:

\[ O_t = \min(C - l_t^r - r_t - l_t^s d_t \phi) \]  \hspace{1cm} (1)

\[ l_t^r + r_t + l_t^s + \gamma_t \leq C \]  \hspace{1cm} (2)

\[ u_t(l_t^r + r_t) + l_t^s + \gamma_t + l_t^s + s_t \leq C, \forall t = 0, \ldots, T - 1 \]  \hspace{1cm} (3)

\[ \pi = \max_{\gamma_t} \sum_{t=0}^{T-1} r_t \phi + \alpha p u_t(l_t^r + r_t) + \beta p(l_t^s + s_t) \]  \hspace{1cm} (4)

where, \( \pi = \) revenue and rest of symbols and definitions of the above equations shown in the following Table 5,

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O_t )</td>
<td>Extra on-demand instance without over capacity</td>
<td>( r_t )</td>
<td>No reservation instances at time ( t )</td>
<td>( o_t )</td>
<td>No of on-demand instance committed at time ( t )</td>
<td>( l_t^s )</td>
<td>No of spot instance at time ( t )</td>
</tr>
<tr>
<td>( C )</td>
<td>The total IaaS data center Capacity</td>
<td>( l_t^p )</td>
<td>No previously committed reserved</td>
<td>( u_t )</td>
<td>The real utilized reserved capacity ratio</td>
<td>( s_t )</td>
<td>No of accepted spot instances at time ( t )</td>
</tr>
<tr>
<td>( l_t^r )</td>
<td>No of committed Reserved Instances</td>
<td>( d_t^o )</td>
<td>Time of instance duration at time ( t )</td>
<td>( \gamma_t )</td>
<td>Predicted duration time of the workload</td>
<td>( t )</td>
<td>Time</td>
</tr>
<tr>
<td>( \phi )</td>
<td>Upfront reserved fee (premium)</td>
<td>( a )</td>
<td>Reserved discount rate related to on-demand price</td>
<td>( p )</td>
<td>Price of On-demand per hour</td>
<td>( \beta )</td>
<td>Ratio of average spot price regard to on-demand</td>
</tr>
</tbody>
</table>

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 schemes: reserved, on-demand and spot. The paper presented a novel idea about how to maximize cloud revenue with the defined or fixed cloud capacity based on different cloud pricing schemes. However, there are some gaps regarding problem assumptions: 1) the revenue function excluded the cost component; 2) There is no one unified price for on-demand instance in reality. It varies from one instance to another. It is dependent on configuration of CPU, GPU, RAM, network bandwidth, storage size, and other cloud features; 3) AWS is charging on hourly base for on-demand instance while Google Cloud Platform (GCP) is charging on minute base; 4) Based on the AWS price scheme, spot instances can be emptied in 2 minutes warming 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 presented the cloud price solution could be considered as a market based pricing in order to maximize CSP 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 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 stochastic optimal policy changes price continuously (or the price change is a continuous variable), then the expected revenue function:

\[ J_u(x, t) = E_u[\int_0^t p(s) dX(s)], \forall t > 0 ; \int_0^t dX(s) \leq x \]  \hspace{1cm} (5)

Subsequently, the problem becomes how to maximize this revenue function

\[ f'(x, t) = \sup_{u \in U} J_u(x, t) \]  \hspace{1cm} (6)
Again, all symbols and definitions of the above equations listed in Table 6.

Table 6. Symbols and Definitions

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_u$</td>
<td>Sales Revenue with $u$ number of price policy</td>
<td>$E_u$</td>
<td>Expected Value from $u$ number of price policy</td>
<td>$p(s)$</td>
<td>Price variable in &quot;s&quot; time interval</td>
<td>$s$</td>
<td>Time variable between [0, t]</td>
</tr>
<tr>
<td>$X(s)$</td>
<td>Number of spot instance sold within time &quot;s&quot;</td>
<td>$x$</td>
<td>The quantity of spot instance within the capacity</td>
<td>$t$</td>
<td>Finite time $t&gt;0$</td>
<td>$\mathcal{U}$</td>
<td>The class of all pricing policies</td>
</tr>
<tr>
<td>$f(p)$</td>
<td>Arrival rate function</td>
<td>$g(p)$</td>
<td>Departure rate function</td>
<td>$\delta t$</td>
<td>Price at a small interval of time</td>
<td>$o(\delta t)$</td>
<td>More than one spot instance sales probability</td>
</tr>
</tbody>
</table>

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 might be validated 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 low prices and low price variance means 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 Fig.9. In one case, the spot price reached a ridiculously high price - $999.00.[58]

Normally, the spot instance price variation 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

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

The model also assumed that all costs for a CSP to operation a cloud environment is irrelevant to its revenue maximization based on Greenberg [60] works. That is to say; there is no marginal cost. This assumption might require further investigation. Moreover, the author’s interpretation of Greenberg’s work needs further consolidation.

The paper also assumed that cloud customers are the price takers. They do not have any influence on the AWS spot instance price. 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 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 what 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 pay $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 reserve price at any time, the old bidder has two minutes warning time to terminate his running instance. In this case, AWS will not charge its customer if the instance running time is less than one hour. Clearly, the existing customer can either revise his/her upper ceiling reserved price or move his/her workload to “on-demand” instance. Sometimes, 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 less than the “on-demand” price. Recently, AWS has capped the highest bidding price four times of “on-demand” price. Moreover, AWS also offers up to 6 hours spot instance to accommodate different types of workloads. These new rules will change the game of VM resource bidding.

![AWS Spot Bid Advisor](image)

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:

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

(9)

where, $\delta_i$ is the narrow band, $a_1$ is the coefficient, $\epsilon(\sigma)$ is the white noise, $p_i$ is a price at any time “i”. It is an empirical observation. If the authors adopt auto-regression or statistical method, their result and conclusion could have more weight when the p-value was demonstrated. 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 “on-demand” price. The authors 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 as:

$$p^* = \max\left\{\pi F^{-1}(1 - \frac{t_k}{t_s})\right\}$$

(10)

For the continuous bidding strategy, the optimal bidding price is

$$p^* = \psi^{-1}\left(\frac{t_k}{t_s} - 1\right), \quad \psi(p) = F_{\pi}(p)\left(\int_p^{\infty} x f_{\pi}(x) dx\right) - 1$$

(11)

Where $\psi^{-1}(\cdot)$ the inverse function of $\psi(p)$

And then, these bid strategies were applied for MapReduce type of workload to obtain the optimal number of slave nodes $M$ running in parallel so that cloud user can minimize spot instance cost.

$$\max_{i=1,\ldots,M} T_i F_{\pi}(p) = \frac{t_s + t_o - M t_r}{M\left(1 - \frac{t_o}{t_k}\left(1 - F_{\pi}(p)\right)\right)}$$

(12)
Zheng’s paper 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.) The paper 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 where the capacity utilization function was coming from: $\beta \log(1 + N(t))$ (where $N(t)$ = the number of accepted bid instance). 4.) The assumption of uniform distribution for bid prices in the section 4.1 appears to be contradicting with the later section 4.3 of bid prices distribution, namely Pareto and exponential distribution. 5.) The paper intended to isolate the issue of the spot price from other prices such as on-demand and reserved but in reality, the CSP has a big cloud resource pool for all price models. It is an ideal assumption to isolate the spot price alone. 6.) The assumption of workload is i.i.d is not very clear. 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 applications or processes (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 interests 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 may be due to the scientific computation (or workload) and a heavy discount in comparison with “on-demand” and reserved price models. The basic idea of spot instance mechanism can be considered as an analogy of a spot pricing of electricity in an energy market [71]. We can categorize these papers into three different focuses: 1) Achieve the best SLA by optimizing the cost for cloud users [63][64][67][69]. 2) Maximize revenue for CSP and Cloud Service Broker (CSB)[55][56][62][72]. 3) Bidding strategies to minimize computational cost for cloud users [66][68][70]. 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, the reality is that AWS can terminate any spot instance at any time although it gives you only 2 minutes warning time. It is very challenging to consider any logic or rational pattern behind AWS to terminate any spot instance. A small SaaS company –MOZ’s experience of 26/Sept/2011 [73] provided a perfect example shown that it would be a very high risk to leverage the lower spot instance price for SLA services delivery. Due to MOZ out of the bid⁹, all MOZ¹⁰ services had been shot down. 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, customer experiences and company’s brand or reputation. That is why MOZ had switched its cloud infrastructure from a public cloud to colocation [74] in 2013. Normally, the spot pricing instance is not suitable for business applications, especially for mission-critical applications. However, many large and medium-sized enterprises or even some small firms require mission-critical IT infrastructure. Spot instance can only be applied to no-SLA and interruptible workloads. However, the majority of computation-intensive workloads, such as batch processing, encoding or decoding, rendering, modeling or continuous integration, cannot generate checkpoints over its multi-hour running period. Perhaps, this is why only AWS offers the spot pricing model. Other CSPs do not offer spot pricing. From this perspective, the spot price is similar to free-trial as one of the marketing campaign strategies. If any business customer would like to solely rely on the spot instance to run a mission-critical application for cost saving, it ought to be failed. There is no free lunch.

---

⁹ MOZ reserved bid was $2 per instance for more than 3 years

¹⁰ MOZ provides Search Engine Optimization (SEO) web crawler services to its customers. MOZ charges its customers on monthly subscription fee.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p^*$</td>
<td>Optimal bid price</td>
<td>$p$</td>
<td>User bid price</td>
<td>$F^{-1}_\pi$</td>
<td>The inverse function of $F_\pi(p)$</td>
<td>$t_r$</td>
<td>Recovery time from an interruption</td>
</tr>
<tr>
<td>$\pi$</td>
<td>Reserved spot price</td>
<td>$\pi(t)$</td>
<td>Market spot instance price at $t$</td>
<td>$t_k$</td>
<td>Length of one-time slot (AWS = 1 hour)</td>
<td>$x$</td>
<td>Bid price variable</td>
</tr>
<tr>
<td>$F_\pi(p)$</td>
<td>the cumulative distribution function of user bid price $p &gt; market price \pi(t)$ at time “t”.</td>
<td>$f_\pi(x)$</td>
<td>the probability distribution of the spot prices, (uniform or Pareto)</td>
<td>$t_s$</td>
<td>Job execution time (without interruptions)</td>
<td>$dx$</td>
<td>A derivative of bid price variable</td>
</tr>
<tr>
<td>$M$</td>
<td>Number of parallel slave nodes</td>
<td>$T_i$</td>
<td>Total time</td>
<td>$i$</td>
<td>Sub-job of MapReduce workload</td>
<td>$r_o$</td>
<td>a constant additional overhead time from splitting the job</td>
</tr>
</tbody>
</table>
### 4.3 Cost-Based Cloud Pricing

As early as 2008, Greenberg et al. [60] discussed the cost based model regarding cloud data centers. It provided a rough estimation of infrastructure cost for cloud services. Some critical assumptions of their estimation were 50,000 physic servers or nodes and 5% of interest rate for capital investment, $3,000 per server, three year lifecycle time and electricity price $0.07/per KWH. The guideline to build own cloud data center showed in Table 8.

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

They highlight major issue across many data centers at that time (before 2008), which has a lower utilization rate of data center resources. The authors 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 network. However, the estimated costs for the cloud data center are dependent on each case and the location of the data center. For example, the authors assumed the electricity price is $0.07/per KWH. This price estimation is at the lower end [75]. The average price of electricity power cross developed nation (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 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.

In comparison with Greenberg’s approximation estimation, Walker [77][78] identified the precise costs of both CPU and storage for Present Net Value (NPV) in comparison with AWS EC2 and S3 (or public cloud).

\[
NPV = \sum_{t=0}^{Y-1} \frac{C_T}{(1+k)^t}, \quad NPC = TC \times \sum_{t=0}^{Y-1} \left(\frac{1}{\sqrt{2}}\right)^t \Rightarrow NPC = TC \times \frac{1 - \left(\frac{1}{\sqrt{2}}\right)^Y}{1 - \frac{1}{\sqrt{2}}}, \quad (13)
\]

\[
NPC = Y \times TC, \quad R = \frac{NPV}{NPC}
\]

\[
PC = \frac{FC}{(\sqrt{2})^t}, \quad TC = TCPU \times H \times \mu, \quad R_p(purchase) = \frac{NPV}{NPC} = \frac{\left(1 - \frac{1}{\sqrt{2}}\right) \times \sum_{t=0}^{Y-1} \frac{C_T}{(1+k)^t}}{1 - \left(\frac{1}{\sqrt{2}}\right)^Y} \times TC
\]

\[
R_{up}(purchase – upgrade) = \frac{C_0 \times \sum_{t=0}^{Y-1} \frac{C_T - A}{(1+k)^t}}{Y \times TC}, \quad R_l(lease) = \frac{\sum_{t=0}^{Y-1} \frac{C_T}{(1+k)^t}}{Y \times TC}
\]

The descriptions of all symbols for the above equations defined in Table 9.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(R_p)</td>
<td>The real cost of CPU hour for purchase case</td>
</tr>
<tr>
<td>(C_0)</td>
<td>Hardware acquisition cost</td>
</tr>
<tr>
<td>(C_T)</td>
<td>Operation cost at each year T</td>
</tr>
<tr>
<td>(TC)</td>
<td>Total Useful Capacity</td>
</tr>
<tr>
<td>(T)</td>
<td>Asset value at Year T</td>
</tr>
<tr>
<td>(A)</td>
<td>The server cluster’s original purchase cost</td>
</tr>
<tr>
<td>(R_{up})</td>
<td>The real cost of purchase and upgrade case</td>
</tr>
<tr>
<td>(NPC)</td>
<td>Net Present Capacity</td>
</tr>
<tr>
<td>(R_l)</td>
<td>The real cost of Lease</td>
</tr>
<tr>
<td>(\mu)</td>
<td>The expected server utilization rate</td>
</tr>
<tr>
<td>(\text{PC})</td>
<td>Present Capacity</td>
</tr>
<tr>
<td>(\text{TCPU})</td>
<td>Total CPU cores of the server cluster</td>
</tr>
<tr>
<td>(\text{Future Capacity})</td>
<td>The expected number of operation hours</td>
</tr>
</tbody>
</table>
According to Walker’s calculations with assumptions of 90% of server utilization rate, 5% capital cost and a cluster or 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. Finally, if the capacity utilization rate is less than 40%, it would always be more reasonable to use the cloud infrastructure.

Based on the same principle (NPV or principle of finance), Walker presented a formula for the enterprise storage cost in the comparison between own build or purchases and public cloud.

\[
\Delta NPV = \sum_{T=0}^{N} \left( C_T - E_T + L_T \right) \frac{S}{(1 + r)^T} + S = R \times \Omega \times [V_T] \Omega \times K \times e^{-0.438T} 
\]

\[
C_T = -r \times H_T - (365 \times 24) \times \delta \times (P_C + P_D \times [V_T] \Omega) 
\]

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

Again, all symbols are defined in the following Table 10.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\delta)</td>
<td>Cost of electric utility ($/kWh)</td>
</tr>
<tr>
<td>(C)</td>
<td>Disk controller unit cost ($)</td>
</tr>
<tr>
<td>(L_T)</td>
<td>Expected annual per-Gbyte lease payment ($/Gbyte/year)</td>
</tr>
<tr>
<td>(\Delta NPV)</td>
<td>Incremental Net Present Value</td>
</tr>
<tr>
<td>(\Omega)</td>
<td>Size of purchased disk drives (Gbytes)</td>
</tr>
<tr>
<td>(H_T)</td>
<td>Annual human operator salary ($)</td>
</tr>
<tr>
<td>(P_C)</td>
<td>Disk controller power consumption (kW)</td>
</tr>
<tr>
<td>(E_T)</td>
<td>the operating cost in year T,</td>
</tr>
<tr>
<td>(\rho)</td>
<td>Proportional difference between human effort in maintaining a purchased versus a leased storage infrastructure</td>
</tr>
<tr>
<td>(I_F)</td>
<td>Risk-free interest rate (%)</td>
</tr>
<tr>
<td>(P_D)</td>
<td>Disk drive power consumption (kW)</td>
</tr>
<tr>
<td>(E_T)</td>
<td>the capital cost in year T</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>Used disk depreciation factor on salvage ([0.0, 1.0])</td>
</tr>
<tr>
<td>(K)</td>
<td>Current per-Gbyte storage price ($/Gbyte)</td>
</tr>
<tr>
<td>(V_T)</td>
<td>Expected storage requirement in year T (Gbytes)</td>
</tr>
<tr>
<td>(S)</td>
<td>Expected end-of-life disk salvage value</td>
</tr>
</tbody>
</table>

The hypothetical assumption for cloud storage pricing was based on the threshold levels of storage size (As shown in Table 11). It means that Cloud Service Providers (CSP) often give a volume discount, which is a kind of linear discount.

| Table 11. Hypothetical Assumption of Cloud Storage Pricing Structure (2010)[78] |
|---------------------|---------------------|---------------------|---------------------|
| 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 be different cost-based pricing model for cloud storage (See Table 12). The price range could be as high as 21 times difference. Moreover, each CSP may give different depreciation rates of cloud storage price each year. This means the \(L_T\) is a time variable, not a constant.

| Table 12. 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 |

The paper’s conclusion was if a decision maker would like to operate the storage system more than four years, the solution of building own storage infrastructure is a preferred option otherwise cloud solution would achieve 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 through a specified period. However, the predicted cost per Gbytes is dependent on previous observation. Different sources of
price data collection could lead to various results. For example, if we adopt McCallum’s dataset [79], the $G_x = 1.3314e^{-0.067}$ (the deprecation rate of $/per GB$) between Apr-2003 and Sep-08 (Refer to Fig. 10, a)

Moreover, if we take the period from 2003 to 2017, the best format to fit historical HDD price data set would be logarithm rather than an exponential one (Refer to Fig.10, 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 Fig. 10, c)

Fig.10: Hard Disk Drive Price

It indicates that $E_T$ 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 cloud be fluctuated along these changes. Walker’s cost-based pricing can be considered as consumption-based or resource-based pricing model as well. Pricing cloud resources is another type of cost-based price. Xu et al. [80] developed a preliminary price model for cloud resources. The essence of this model is an Isoelastic or constant elasticity function (a particular case = constant relative risk aversion (CARRA) of Hyperbolic Absolute Risk Aversion (HARA)) based on economic utility theory [81]. It means CSP can maximize revenue if tenants or cloud consumers will make rational choices under risk aversion to maximize the expected value of a concave Von Neumann-Morgenstern utility function. CSP’s revenue maximization strategy is presented as essential price discrimination, throttling, SLA performance functions that are defined as follows:

$$vU(x), \quad \max_{x} vU(x) - px, \quad U(x) = \begin{cases} \frac{x^{1-\alpha}}{1-\alpha}, & \alpha \neq 1 \\ \ln(x), & \alpha = 1 \end{cases}$$

(19)

$$U(x) = (1 - \alpha)^{-1}x^{1-\alpha} = \frac{x^{1-\alpha}}{(1 - \alpha)}, \quad \alpha \in (0, 1), \quad x = D_v(p) = \left(\frac{v}{p}\right)^\frac{1}{\alpha}$$

(20)

$$E_d = -\frac{dD_v}{dp} \cdot \frac{p}{D_v} = \frac{1}{\alpha} \quad R_v(p) = pD_v(p) = \frac{v}{p^\frac{1}{\alpha}}$$

(21)

**Basic Choice**

$$\max_{p^v} \int_{v_0}^{v_1} R_v(p)f(v)dv, \quad s.t. \int_{v_0}^{v_1} D_v(p)f(v)dv \leq C$$

(22)

$$S_v(p) = vU[D_v(p)] - pD_v(p) = \frac{ap}{1 - \alpha} \left(\frac{v}{p}\right)^\frac{1}{\alpha}, \quad S_v(p) \geq 0, \forall v, \quad Over \ p$$

(23)

**1st Order Price Discrimination Choice**

$$\max_{p^v} \int_{v_0}^{v_1} R_v(p)f(v)dv, \quad s.t. \int_{v_0}^{v_1} D_v(p)f(v)dv \leq C, \quad S_v(p) \geq 0, \forall v, \quad Over \ (p_v)$$

$$p^* = \left(\frac{B}{C}\right)^\alpha = p_c, \quad R^* = B^\alpha C^{1-\alpha}, \quad S^* = \frac{\alpha}{1 - \alpha} B^\alpha C^{1-\alpha}, \quad B = \int_{v_0}^{v_1} v^\alpha f(v)dv$$

(24)

**Throttling Option Choice**

$$\max_{p^v} \int_{v_0}^{v_1} R_v(p)f(v)dv, \quad s.t. \int_{v_0}^{v_1} \beta D_v(p)f(v)dv \leq C, \quad S_v(p, \beta) = p \left(\frac{v}{p}\right)^\frac{1}{\alpha} \left(\frac{\beta^{1-\alpha} - 1}{1 - \alpha}\right) \geq 0, \quad over \ p, \beta$$

(25)

**Performance or SLA Guarantees Choice**

$$\max_{p^v} \int_{v_0}^{v_1} R_v(p,q)f(v)dv, \quad s.t. \int_{v_0}^{v_1} D_v(p)f(v)dv \leq C$$

$$S_v(p, q) = vU[D_v(p)] - pD_v(p) - q = \frac{ap}{1 - \alpha} \left(\frac{v}{p}\right)^\frac{1}{\alpha} \geq 0, \forall v, \quad Over \ p, q$$

(27)

Profit Max- Operation Cost and Capacity Right Sizing
\[
\max_{p>0} \int_{v_0}^{v_1} (R_v(p) - E_v(p)) f(v) dv, \quad \text{s.t.} \int_{v_0}^{v_1} D_v(p) f(v) dv \leq C
\]  

\[E(x) = e_0 + e_1 x, \quad \text{if} \ x = D_v(p), \quad E_v(p) = e_0 + e_1 \left(\frac{v}{p}\right)^\frac{1}{\gamma}, \quad S_v(p) \geq 0, \forall v, \over p\]

All the above symbols and their definitions are shown in Table 13.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(v)</td>
<td>utility level</td>
<td>(x = D_v(p))</td>
<td>VM or instance demand</td>
<td>(S_v(p))</td>
<td>Tenant’s surplus</td>
<td>(q)</td>
<td>Extra charge for SLA performance</td>
</tr>
<tr>
<td>(U(x))</td>
<td>utility function</td>
<td>(E_d)</td>
<td>Price elasticity</td>
<td>(C)</td>
<td>The capacity of Cloud instance</td>
<td>(e_0)</td>
<td>Fixed energy costs independent of workload</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>The inverse of Price elasticity</td>
<td>(R_v(p))</td>
<td>CSP’s Revenue</td>
<td>(\beta)</td>
<td>A fraction of throttling</td>
<td>(e_1)</td>
<td>Variable energy costs per unit of resources</td>
</tr>
<tr>
<td>(p)</td>
<td>Virtual machine or cloud instance price</td>
<td>(f(v))</td>
<td>the density function of probabilistic utility level distribution</td>
<td>(B)</td>
<td>Constant to optimize price (p) for maximizing the CSP revenue</td>
<td>(E_v(p))</td>
<td>Major operational cost</td>
</tr>
</tbody>
</table>

The main contribution of the Xu’s paper is that it adopted the iso-elasticity function for the utility. It emphasized that CSP could leverage cloud customers’ surplus to maximize its revenue. However, this pricing model for cloud resource has a few issues to be implemented in practice. 1.) The assumption of CSP (or IaaS) that has a monopoly market power is fine for new cloud service feature but it might need further investigation to demonstrate that is a case for a commodity type of IaaS because we know that there are numerous CSP in the IaaS market, such as GCP, MS Azure, Rackspace, IBM Softlayer, AWS, Dimension Data, Oracle Cloud, and HP. 2.) Intuitively, it is challenging to charge cloud customers 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 group) pricing discrimination. 3.) The assumption of throttling is unclear. Due to the nature of online price, CSP has to specify its performance of VM or instance. If a CSP reduces the specified VM performance (such as CPU speed, RAM, and storage size), this means it cannot fulfill its legal contract. 4.) Ultimately, adopting the isoleastic utility function to model cloud consumers’ utility is still opened to further discussion because practically, if the quantity of cloud demand is increasing, the price elasticity would be quite challenging to be kept the same unless increase cloud capacity. 5.) Definition of probability density function \(f(v)\) appears to be too arbitrary. In addition, Google Cloud Platform (GCP) and the Amazon Web Service (AWS) price scheme are different. GCP price calculation for on-demand instance is based on the first 10 minutes + 1 minute rather than per hour base as AWS.

The author assumed the elasticity that is equal to \(E_d = \frac{1}{a} = 3\) seems to be lack of enough support evidences. Subsequently, the level of utility \(v = p^{\frac{1}{a}}\). If we use the paper’s price assumption: \(p = 0.08/\text{per hour for a small Linux instance}\), then utility level \(v = 0.08^{\frac{1}{3}}\). And then the paper used Google, RICC and ANL cluster trace information to consolidate the utility density distribution. Based on the Alam et al [82] research work, the workload pattern of Google cluster trace is more like trinodal the pattern and long jobs would consume more resource at all time. 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 commercial cloud resources modelling. Finally, the revenue score only measures a firm’s sale. Although the sale enhancement could be one of the goals for any commercial CSP, the ultimate goal is the profit. The revenue maximization is not equal to profit maximization. Sometime, it might mean losing money if the instance cost is higher than the sales price, which the higher revenue, the large deficit is. So, without considering instance cost, the cloud revenue maximization appears to be premature. According to Belleflamme and Pietz [85], the above revenue maximization function (monopoly pricing formula) should be altered as:

\[
\max_{D_v} \pi(D_v) = D_v p(D_v) - C(D_v)
\]  

where \(C\) is the average cost and both price \(P\) and Cost \(C\) are the functions of demand: \(D_v\), where \(D_v p(D_v) = R_v(p)\).

So far, we only focus on the cost part of cloud pricing, but this part pricing is playing a partial role in the principle of pricing modeling. The crucial concept of price modeling is not about pricing physical hardware but cloud customer’s utility. It is an abstract measurement of how much is worth to a particular cloud consumer. The utility is also an intangible measurement of satisfaction in term of individual cloud consumer’s values to specific cloud service. So, the ultimate measurement of cloud price is value.

### 4.4 Value-Based Cloud Pricing

For the value-based cloud pricing model, one of the scientific approaches is a so-called hedonic method. It has been widely applied to the consumer price index (CPI) by many OECD countries, such as US Bureau of Labor Statistics (BLS), Australia Bureau of Statistics (ABS), British Office for National Statistics (ONS), Germany Federal Statistical Office (Destatis), etc.
El Kihal et al. [84], Weinman [4], Mitropoulou [85] and Zhang [86] either proposed or presented a hedonic method for cloud pricing. El Kihal showed the comparison results among major CSPs (AWS, IBM Cloud, Microsoft Azure, Terremark, and Google App Engine) in term of three cloud characteristics: memory ($ per GB), CPU ($ per CPU) and Storage ($ per 100GB). Overall, the paper had a gap to explain the details that how the dataset was collected and how many cloud instances were gathered.

$$BA_p = \beta_{0p} + \sum_{i=1}^{l} \beta_{ip}x_i + \epsilon_p, \{p \in P\}$$  \hspace{1cm} (32)

All symbols are defined in Table 14.

**Table 14. Linear Hedonic Regression Model**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$BA_p$</td>
<td>Price plan or billing amount</td>
<td>$\beta_{0p}$</td>
<td>the constant coefficient of linear regression</td>
<td>$x_i$</td>
<td>Hedonic Characteristics</td>
<td>$\epsilon_p$</td>
<td>Error term of the regression equation</td>
</tr>
<tr>
<td>$p$</td>
<td>Number of CSP</td>
<td>$\beta_{ip}$</td>
<td>Parameters of hedonic characteristics</td>
<td></td>
<td>Number of hedonic characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total # of Characteristics</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The result experiment indicated that the 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 result seems to be inaccurate. 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 issue. 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.”

In comparison with El Kihal et al.’s [84] paper, Mitropoulou et al. [85] made some progress of the hedonic method. The article 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 article is that if the paper adopted a hedonic index measurement, it needs a base period for comparison.

This issue was picked up by Zhang’s thesis [86]. Based on Pakes [87]’s seminal work, the author explained the fundamental concept of the hedonic method based on the underlying assumption of economic theory. It is the expected values of cloud service characteristics. The primary contribution of the paper was to introduce the time dummy variable for the hedonic model of cloud price to analysis AWS’ cross-sectional data between 2009 and 2015.

$$h_{DV}(x_i); \ln P_{i,t} = \alpha + \sum_{k} \beta_k x_{k,i,t} + \sum_{t} \delta_t D_{i,t} + \epsilon_{i,t}$$ \hspace{1cm} (33)

where the meanings of all symbols for the above equations are shown in Table 15.

**Table 15. Hedonic Pricing Model with Time Dummy Variables**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_{DV}(x_i)$</td>
<td>Hedonic values with a dummy variable</td>
<td>$x_i$</td>
<td>Characteristics of Cloud services</td>
<td>$(x_{-i},P_{-i})$</td>
<td>Characteristics and price of other goods</td>
<td>$K$</td>
<td>Number of characteristics</td>
</tr>
<tr>
<td>$P_{i,t}$</td>
<td>Price of goods “i” at time</td>
<td>$D_{i,t}$</td>
<td>Dummy Variable</td>
<td>$\delta_t$</td>
<td>Time dummy coefficients</td>
<td></td>
<td>Number of the time period (year)</td>
</tr>
<tr>
<td>$mC()$</td>
<td>Marginal cost</td>
<td>$\alpha$</td>
<td>Constant value</td>
<td>$X_{k,i,t}$</td>
<td>The vector of cloud characteristics</td>
<td></td>
<td>Consumer preferences over characteristics</td>
</tr>
<tr>
<td>$D_{i,()}$</td>
<td>The demand function for good i</td>
<td>$\beta_k$</td>
<td>Coefficient of characteristics</td>
<td>$\epsilon_{i,t}$</td>
<td>Error term</td>
<td>$A$</td>
<td>Consumer preferences over characteristics</td>
</tr>
</tbody>
</table>

Based on the experiment results, the adjusted R2 value was 0.9792 for 277 data points. In comparison with other papers, his 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. Furthermore, the p-value of storage is less significant than other cloud service characteristics. The value of the storage coefficient showed as a negative.

As the author concluded, the major issues of the paper are 1.) A small sample of data, and 2.) the hedonic method that some hidden cloud characteristics were left out. 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. Performance-based pricing model [89] is associated with cloud resource and applications risks. Feature-based pricing[90] that is related to prioritizing cloud features. Service-based pricing model [91] correlates to the Service Level Agreement (SLA).

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. Naturally, all customers would like to have free or near-free cloud resource 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 cost under certain assumptions. The pricing scheme 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 reliant 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 claimed that it is an integrated approach to price IaaS resource 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 it might 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 SLA features are quite challenging to be measured by the cost-based price. They are built into cloud service as a whole for a particular CPS to differentiate its service from other CPSs.

Despite the traditional value-based pricing strategy, AWS first launched the new value-based pricing model in 2014, namely Lambda function. It is delivered by the serverless sandbox technology, which is also called Function as a Service (FaaS) delivery model. It is supported by the Docker container or 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 Sbarsky showed [95] the details how to calculate the total cost of AWS Lambda function. We can use the following equation 34 for the calculation of AWS Lambda price.

$$P_t = h_r + m_r = (a[X_{100}]h - k) \times h + \left(\frac{a[X_{100}]h R}{10} - g\right) \times r_m \quad (34)$$

where, $P_t$ is the total price of the Lambda function per month, $h_r$ is the hit price and $m_r$ is a resource memory price. $a$ is the constant value of second per month = 2,628,000. $[X_{100}]$ is the round up integer of code execution time per 100ms. $h$ is the hit rate per second (execution rate of code request). If the user’s code execution time is less than 100 ms, for example, $C_{100} = 85ms$, “$[X_{100}]$” is equal to or normalized to 1, if the code’s execution time is more than 100, for example, $X_{100} = 101$, “$[X_{100}]$” is equal to 2. $R$ is the allocated memory resource, e.g., 256 MB-second. $y$ is the baseline memory 1024 MB-second (reference price). $r_m$ is the price rate $\.020/per million hits (Lambda@Edge). If the code execution time/per 100ms, it would be free for compute resource. However, Lambda@Edge has no free allowance. $g$ is the free allowance of 1024 MB-s is 00, 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 256 MB/per 100ms, we should have $h = 50, [X_{100}] = 2, y = 1042 MB/per 100ms, The total monthly bill is: $P_t = h_r + m_r = (a[X_{100}]h - k) \times h + \left(\frac{a[X_{100}]h R}{10} - g\right) \times r_m = (2,628,000 \times [125_{100}] \times 50 - 1,000,000) \times 0.00000002 + \left(\frac{2,628,000 \times [125_{100}] \times 50}{10} - 400,000\right) \times 0.0000001667 = $52.36 + $10.95 = $63.31/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) compute execution time for another user. On the other hand, the cloud consumers only pay what the code execution time or the usage is if this execution period is very close but less than the time of 100ms. Obviously, this price does not include the cost of storage, API gateway, and data egress.

The bad news for this model is if the number of hits/per second is remarkably higher, the cost of Code of Demand (CoD) could be out of a hand. Sometimes, it could be three times higher than VM on-demand [93].

The value-based pricing models are often confused with cost-based pricing models as both pricing models include cost or physical components. However, value-based pricing model can extract the relationship or instrumental value, which is also called as extrinsic values (such as burstable CPU, auto-scaling, Open API, serverless platform, etc.), from its intrinsic values (or non-relationship values that can only reflect its own value, such as CPU, RAM, network bandwidth). With a statistical solution, the hedonic pricing model can extract the relationship values from intrinsic variables. Likewise, the pricing model of AWS Lambda function is to abstract away from the cloud physical cloud resources further. AWS only charges the cloud service function based on the code execution time and the number of hits in term of a relationship of memory, CPU, storage, and network outbound size or resource.

4.5 Summary

We have walked through the papers of cloud pricing model from 2008 to 2016. Notably, we have a more in-depth analysis of 10 papers. These papers can be summarized in three basic pricing strategies according to the principle of axiology. We summarize these cloud pricing models in Table 16. Chronologically, the cloud price modeling method has been leaning towards value-based pricing.
Table 16. Summary of Cloud Pricing Models Survey

<table>
<thead>
<tr>
<th>Category of pricing models</th>
<th>Mathematical Equation of model</th>
<th>Main Contributions</th>
<th>Potential Issues and Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marketed Based pricing: Toosi et al.’s Max CSP Revenue (2014)</td>
<td>$\pi = \max_{i=1, \ldots, M} T_F(p)$</td>
<td>The main contributions of this paper offer an alternative pricing model for CSP to price its spot instance dynamically (if the spot instance is not auction base).</td>
<td>However, the spot pricing cannot be generalized to all instances. In one case, the spot price reached a ridiculously high price - $999.00. Usually, the spot instance price variation with time is neither convex nor continuous. Two critical functions are defined as more like a power function rather than a Poisson distribution function.</td>
</tr>
<tr>
<td>Marketed Based pricing: Xu et al.’s dynamic pricing model (2013)</td>
<td>$f'(x, t) = \sup_{u \in U} \left[ p(x) dX(s) \right] \forall t &gt; 0$</td>
<td>It intended to unveil the spot price mechanism of AWS.</td>
<td>If the authors adopt the auto-regression or statistical method, the result and conclusion may have more weight when the p-value is demonstrated.</td>
</tr>
<tr>
<td>Marketed Based pricing: Orna, et al. Traceable data (2013)</td>
<td>$\delta_i = -a_i \delta_{i-1} + \epsilon(\sigma)$, and $p_i = p_{i-1} + \delta_i$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marketed Based Pricing: Zheng, Liang et al. (2015)</td>
<td>$R_p(p) = \frac{NPV}{NPC} = \frac{1 - \frac{1}{\sqrt{2}}} {\sum (1 + k)^{1/2} \times TC \times \sum (1 + k)^{1/2} \times TC}$</td>
<td>The paper highlighted a significant issue across many data centers at that time (before 2008). It provided a rough estimation of cloud data center element cost.</td>
<td>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.</td>
</tr>
<tr>
<td>Cost-Based Pricing: Greenberg et al. (2008)</td>
<td>$R_p(p) = \frac{NPV}{NPC} = \frac{1 - \frac{1}{\sqrt{2}}} {\sum (1 + k)^{1/2} \times TC \times \sum (1 + k)^{1/2} \times TC}$</td>
<td>The interpretation of their experiment result seems to be inaccurate. The adjusted R square value is only between 43%~69%. It means the linear regression is not “goodness of fit.”</td>
<td></td>
</tr>
<tr>
<td>Cost-Based Pricing: Walker, Edward (2009)</td>
<td>$\Delta NPV = \sum_0^{T} \frac{C_T}{(1 + l_T)^{T}} + \frac{S}{(1 + l_T)^{T}} - C$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost-Based Pricing: Walker, Edward (2010)</td>
<td>$\max_0^{v_0} \int_{v_0}^{v_1} \left( D_v(p) f(v) dv \right. \left. \right) s.t. \int_{v_0}^{v_1} \left( D_v(p) f(v) dv \right) \leq C$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost-Based Pricing: Xu, Hong, and Baochun Li (2013)</td>
<td>$B_{A}\beta = \beta_{ap} + \sum_{i \in I} \beta_{ap} x_i + \epsilon_p, [p \in P]$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| | | | |
We observe that the commonality of the majority of cloud pricing models is to maximize CSP’s cloud business profits while to minimize the cloud resources consumption. The common trait of existing solutions is the infrastructure cost has a major impact on the pricing model. The differences are value-based pricing driven from the demand side while cost-based pricing is oriented by the supply side. And the market-based pricing takes a balanced view between supply and demand.

During the early era of cloud computing, many researchers mainly focused on the cost issue of cloud infrastructure. Walker’s two papers and Greenberg’s paper provided a typical example. When the cloud market become mature, especially, since AWS launched the spot-instance in 2006, the research topic of cloud pricing model was oriented by market-based pricing. Xu, Orna, and Toosi included on-demand, reserved and spot-instance price models into their solution of profit maximization. Just recently, El Kihal, Mitropoulou, and Zhang adopted the hedonic method, which is considered as value-based pricing to explain the cost mechanism of cloud infrastructure based on both panel and cross-section data. Therefore, there is a knowledge gap for the value-based pricing, which is how to leverage the hedonic approach to extract extrinsic or instrumental values from IaaS.

5. Conclusions and Future Directions

Cloud pricing is moving further away from a physical box oriented model to a virtual machine based model, and then to an abstract sandbox based model. Many CSPs start to offer cloud pricing based on an abstract layer of cloud resource. To some extent, pricing of the serverless sandbox can be considered as modeling No Operation Systems (No OS or NoOps), which is a evolutionary direction from a pure development environment to an integrated environment of both development and operation or DevOps.

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 CSP’s responsibility. They become a part of SLA measurement or service-based pricing. Cloud customers do not have to get their hand 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 the essential tool behind the new 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 considered as online retail-oriented pricing for its cloud services. Azure is software-oriented pricing, and GCP is search engine optimization (SEO) oriented pricing. The other upcoming CSP competitors can leverage different applications strength for their cloud pricing, such as e-commerce, utility services, healthcare, cyber-security, Supervisory Control and Data Acquisition (SCADA), Internet of Things, and Business Intelligence Analytics.

Overall, the cloud computing technologies and cloud pricing strategies have four development trends, which computation resource has moved from statefulness to stateless, IT infrastructure has been transferred from dedicated to the shared base, the software has been gradually shifted from mutability to immutability, and cloud pricing models are moving from cost-based to value-based pricing strategy. These trends are leading towards a hyper-converted architecture or a resource pool of CPU, cache memory, RAM, disk, and network that cloud service are delivered by its extrinsic values rather than pure intrinsic values only. We can summarize these trends in Fig. 11 from three perspectives:

- From a resource perspective, there is a trend towards a stateless architecture that is to enable customers to scale a resource pool quickly. There are two meanings of stateless. One is “thin server with the thick client,” which a server does not have 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 state of a server that traditionally needs defined memory, network bandwidth, storage, and operating system. In a simple term, cloud customer does not need a specified or fixed server box, whether it is 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 resource, the ephemeral nature of computational requirement for server resource is also considered as serverless.
- From an infrastructure perspective, there is a trend of sharing, which aims to maximize the utilization of cloud infrastructure. Up to now, all cloud resources, whether it is built by CSPs or large enterprises, are supported by the physical data centers and communication networks. If the running business applications require mission-critical infrastructure that satisfies peak demands,
the amount of both capital and operational expenditures are significant. Moreover, the proportion of the data center assets value is depreciated sharply due to Moore’s law. Subsequently, sharing infrastructure is an inevitable step to improve the utilization rate of cloud hardware. It also underpins many CSPs’ to transform their pricing strategy from cost-based pricing strategy to value-based pricing. (The key difference between resource and infrastructure is that cloud resource is measured by time while cloud infrastructure refers to a physical object)

- From a software perspective, there is a trend of the distributed system for immutable objects. Traditionally, 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 rather than by multiple times on the top of older binaries. In comparison, immutable software is a new object. A direct replacement does the way for the incremental upgrade. For example, if an old container server needs to upgrade, the fresh new container image will be built, and the old one can be thrown away, and 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.

All the above three cloud developments have set up a cornerstone of the transformation from cost-based pricing to value-based pricing strategy. This transformation does not only emphasize the value of hardware resources, such as RAM size, processor speed, storage capacity, and network performance, but also focus on the value of business application, such as, efficiency of software execution code on demand (CoD), software functionality, speed to market, and a scalability. 451 Research estimated that the market revenue of Docker would grow more than 35 folds from $761m in 2016 to $27 billion in 2020.[96] (as shown in Fig. 11).

![Cost Based Pricing](image1)

**Fig. 11: Future Trends in Cloud Technologies and Cloud Pricing Strategy**

We observe that the overall direction of cloud service is moving from intrinsic value-based into extrinsic value-based pricing. Technologically, this new pricing strategy is supported by serverless, Docker container, Open API, and microservices. From CSP perspective, the implication of this development means it can utilize cloud resource at a granularity level. From a cloud consumer perspective, it means free from locked-in with a particular cloud vendor or CSP. By this evolutionary direction of cloud services, we identify four basic challenges for the future cloud price model development:

- How to drive cost-based pricing to value-based pricing orientation from a business customer perspective.
- How to price the cloud infrastructure resource from statefulness to stateless.
- How to price from mutable to immutable.
- How to price from intrinsic to extrinsic cloud features of a lifecycle.
Nagle’s seminal book [13] provides some clues to deal with these challenges of cloud price modeling. One of the proposals is to establish a value-based metrics from cloud business customer’s perspective, which is to create proactive pricing strategy to understand how, and when to satisfy customers’ application needs and expectation while a CSP can maximize its profitability.

6. References

cGraw
cGraw
cGraw